L-Identification for Sources

Dissertation
zur Erlangung des Doktorgrades
der Fakultät für Mathematik
der Universität Bielefeld

vorgelegt von
Christian Heup

November 2006

Acknowledgement

I would like to thank Prof. Dr. Rudolf Ahlswede for leading my interest to Information Theory and especially to identification. It was a great experience having shared in his wealth of knowledge. I also thank Dr. Ferdinando Cicalese for the helpful discussions and all his efforts. Further, I thank Prof. Dr. Volker Strehl for the fruitful communication which connected identification to harmonic numbers.

I greatly thank my parents for their help and support and Andrea, who I love, for being the most important part of my life.

1st advisor: Dr. Ferdinando Cicalese
2nd advisor: Prof. Dr. Dr. h.c. Rudolf Ahlswede

Gedruckt auf alterungsbeständigem Papier °° ISO 9706

Contents

Introduction

In [2] (Section 2) Ahlswede introduced "a general communication model for one sender". Suppose we have a message set $\mathcal{M} = \{1, ..., M\}$ whose elements are encoded in such a way that information about them can be transmitted over a channel. If this channel is noiseless, i.e. there occur no errors during the transmission, we speak of (noiseless) source coding. In this case it is common to omit the presence of a channel and speak simply of source coding.

What do we mean by information? In Shannon's classical information transmission problem ([15]) the decoder is interested in the message which has been encoded by the encoder. However, the decoder may have different goals. In [2] Ahlswede writes:

"A nice class of such situations can, abstractly, be described by a family $\Pi(\mathcal{M})$ of partitions of $\mathcal{M}$. Decoder $\pi \in \Pi(\mathcal{M})$ wants to know only which member of the partition $\pi = (A_1, ..., A_r)$ contains m, the true message, which is known to the encoder."

In the above citation every partition $\pi \in \Pi(\mathcal{M})$ is identified with a different decoder. Moreover, the author describes some "seemingly natural families of partitions". We focus on the first three models which highlight the differences between classical information transmission and identification. These are

Model 1: $\Pi_S = \{\pi_{sh}\}$, $\pi_{sh} = \{\{m\} \ : \ m \in \mathcal{M}\}$.

Model 2: $\Pi_I = \{\pi_m \ : \ m \in \mathcal{M}\}$, $\pi_m = \{\{m\}, \mathcal{M}\backslash\{m\}\}$.

Model 3: $\Pi_K = \{\pi_S \ : \ |S| = K, S \subset \mathcal{M}\}$, $\pi_S = \{S, \mathcal{M}\backslash S\}$.

The first model describes Shannon's classical transmission problem. Here the decoder wants to know which message has been encoded by the encoder. Let us assume we are given a probability distribution P on the message set. In source coding we consider a *source code* $\mathcal{C} : \mathcal{M} \to \mathcal{Q}^*$. Here $\mathcal{Q}$ is the *q-ary alphabet* $\{0, 1, ..., q-1\}$ and $\mathcal{Q}^* = \bigcup_{d=0}^{\infty} \mathcal{Q}^d$. $\mathcal{C}(m)$ is called the *codeword* of the message m. We further assume that $\mathcal{C}$ is a *prefix code*. That is, no codeword is the prefix of another codeword. The goal of source coding is to construct prefix codes which have a small average codeword length. In other words, the mean of the

codeword lengths should be as small as possible. It is well-known that this value is lower bounded by Shannon's classical entropy

$$H_q(P) = -\sum_{m=1}^{M} p_m \log_q p_m.$$

There exist codes, e.g. Huffman codes ([12]) and Shannon-Fano codes ([10]), which yield an average codeword length of at most $H_q(P) + 1$. The uniform distribution maximizes $H_q(P)$ and it holds that $H_q(1/M, ..., 1/M) = \log_q M$.

In the second model the decoder π_m wants to know whether m occurred or not. This is the identification problem introduced for noisy channel coding in [6] and analyzed inter alia in [7], [11] and [13]. Identification source coding was introduced in [2], continued in [4] and led to the identification entropy ([3])

$$H_{I,q}(P) = \frac{q}{q-1}\left(1 - \sum_{m=1}^{M} p_m^2\right).$$

This entropy function again is maximized by the uniform distribution. Unlike Shannon's entropy it does not grow logarithmically in M but tends to $q/(q-1)$ as M goes to infinity.

A generalization of the identification problem is model 3, which is called *K-identification*. This case arises in several situations. Ahlswede writes: "For instance every person π_S may have a set S of K closest friends and the sender knows that one person $m \in \mathcal{M}$ is sick. All persons π_S want to know whether one of their friends is sick."

Another natural problem is somewhat like the opposite of K-identification. For example, the encoder knows L persons $m_1, ..., m_L \in \mathcal{M}$, who have won a lottery. Every participant, a member of $\mathcal{M}$, wants to know whether or not he or she is among the winners. However, the information in which a participant is interested can no longer be represented by a partition of $\mathcal{M}$. We have to partition $\binom{\mathcal{M}}{L}$ and get

$$\Pi_{L,\text{set}} = \{\pi_m \ : \ m \in \mathcal{M}\}, \tag{0.1}$$

where $\pi_m = \{S_m, \binom{\mathcal{M}}{L} \backslash S_m\}$ and $S_m = \{S \in \binom{\mathcal{M}}{L} \ : \ m \in S\}$. We call this model *$L$-identification for sets*.

One could also think of situations where the L objects, which are known to the encoder, need not be pairwise different. We call this *L-identification for vectors*. The model for this is

$$\Pi_L = \{\pi_m \ : \ m \in \mathcal{M}\}, \tag{0.2}$$

where $\pi_m = \{A_m, \mathcal{M}^L \backslash A_m\}$ and

$$A_m = \{A \in \mathcal{M}^L \ : \ A \text{ has at least one component equal to } m\}.$$

This can also be applied to K-identification so that we obtain

Model 3': $\Pi_{K,\text{vec}} = \{\pi_A \ : \ A = (a_1, ..., a_K) \in \mathcal{M}^K\}$, with

$$\pi_A = \left\{ \bigcup_{i=1}^{K} a_i, \mathcal{M} \backslash \bigcup_{i=1}^{K} \{a_i\} \right\}.$$

This is called *K-identification for vectors* and model 3 *K-identification (for sets).*

The goal of this thesis is the analysis of L-identification in the case of noiseless coding. We call it *L-identification for sources.* However, the concept of L-identification may also be considered in the case of noisy coding. Moreover, we mainly focus on L-identification for vectors. Thus, if we speak in the remainder of L-identification, we shall always mean L-identification for vectors.

The first section provides basic definitions and notation. In Subsection 1.1 we give a short introduction into source coding. A *discrete source* is a pair $(\mathcal{U}, P)$, where the *output space* $\mathcal{U}$ is a finite set of cardinality N and P is a probability distribution on $\mathcal{U}$. Further, a *discrete memoryless source* is a pair $(\mathcal{U}^n, P^n)$, where $\mathcal{U}^n$ is the cartesian product of a finite set $\mathcal{U}$. P^n is a probability distribution on $\mathcal{U}^n$, where the probability of an element $u^n \in \mathcal{U}^n$ is product of the probabilities of its individual components. We further explain what we mean by the *code tree* $T_\mathcal{C}$, which corresponds to a given source code $\mathcal{C}$, and provide some notation.

In Subsection 1.2 we formally define L-identification for sources. Let $L \in \mathbb{N}$ and $(\mathcal{U}^L, P^L)$ be a discrete memoryless source. Due to external constraints (e.g. hardware limitations) all possible *outputs* $u^L = (u_1, ..., u_L) \in \mathcal{U}^L$ have to be encoded. This is done by a *q-ary source code* $\mathcal{C}$ on $\mathcal{U}$. That is, every component u_i of u^L is encoded separately.

Following the model in Equation (0.2) the goal of L-identification is that every *user* $v \in \mathcal{U}$ shall be able to distinguish whether or not he or she occurs at least once as a component of the output vector u_L. Therefore, we encode all users with the same source code $\mathcal{C}$ and compare sequentially the q-bits of the codeword c_v of the user v with the individual q-bits of the codewords $c_{u_1}, ..., c_{u_L}$ of the components of u^L. After every comparison we delete all output components, whose codewords did not coincide during this step with the codeword c_v, from the set of possible candidates. If after some steps all codewords have

been eliminated, the L-identification process terminates with a negative answer. Otherwise we go on until the last q-bit of c_v. The L-identification process terminates with a positive answer if after this last comparison there still are possible candidates left.

The *running time* of q-ary L-identification for given output vector u^L and user v with respect to some code $\mathcal{C}$ is defined as the number of steps until the L-identification process terminates. Since we are given a probability distribution P^L on $\mathcal{U}^L$, we can calculate the mean of the L-identification running time. We call it the *average running time*.

We are interested in several behaviors of the average running time. The first is the *worst-case (average) running time* where we maximize the average running time over all users $v \in \mathcal{U}$. Suppose we have given another probability distribution Q on the set of users $\mathcal{U}$. In this case we calculate the mean of the average running time. This is called the *expected (average) running time*. A special case of this is when $Q = P$. Then we speak of the *symmetric (average) running time*.

We note that the above approach to analyze L-identification can also be used for noiseless K-identification. The only difference between the two models is on which side the L (resp. K) objects are. For L-identification they are on the side of the encoder and for K-identification they are on the side of the decoder. Thus, an immediate conclusion is that the symmetric running time of L- and K-identification is the same if $L = K$. In case of the expected running time we also would have to exchange the probability distributions P and Q. For the worst-case running time such a direct connection has still to be proven.

We begin our analysis of L-identification in Section 2 with two new results for the case $L = 1$. This corresponds to identification for sources, which was introduced before. During this thesis we refer to *(1-)identification for sources* if we speak of identification for sources in order to indicate that identification is a special case of L-identification.

The first result in Subsection 2.1 concerns the case when the q-ary source code $\mathcal{C}$ is a *saturated block code*. This means that all codewords have the same length n and the number of elements equals q^n. We show that for such codes the uniform distribution is optimal for the symmetric running time of (1-)identification. The main part of this subsection is Lemma 2.1 where we provide a modification for a given probability distribution. If this modification is applied iteratively, it results in the uniform distribution and does not increase the symmetric running time of (1-)identification. This result is used by the authors of [5] in the proof of their Lemma 3.

The authors of [4] proved in Theorem 3 that the worst-case running time of binary (1-)identification can be upper bounded by 3 no matter of how big the output space $\mathcal{U}$ is. This was done by an inductive code construction. We show

in Subsection 2.2 how this upper bound can be improved by a slight change of their code construction.

In Section 3 we analyze the asymptotic behavior of the symmetric running time of L-identification for the case that P is the uniform distribution. For this we consider the so-called *balanced Huffman codes for the uniform distribution*. These codes are special cases of the well-known Huffman codes and were introduced in [3].

In Subsection 3.1 we point out an interesting connection between balanced Huffman trees and the colexicographic order. This order can be used to construct a balanced Huffman code.

In Subsection 3.2 we provide Theorem 3.4, the main result of this section. We prove that if we use balanced Huffman codes for the uniform distribution, the symmetric running time of q-ary L-identification asymptotically equals a rational number $K_{L,q}$, which grows logarithmically in L. In fact, we show that this number is an approximation of the L-th harmonic number.

The main result of this thesis is in Section 4 the discovery of the q-ary identification entropy of second degree. We begin this section with the illustration of our approach in finding this entropy function. In order to find a lower bound for 2-identification concerning general distributions we want to apply our asymptotic result of Subsection 3.2 concerning the uniform distribution. Therefore we first establish a connection between 2-identification inside a given code $\mathcal{C}$ and 2-identification inside the concatenated code $\mathcal{C}^n$. It turns out that not only 2-identification comes into play here but also (1-)identification. In the next step we prove that if n is sufficiently large, 2-identification inside the concatenated code can be lower bounded by 2-identification inside a saturated block code of some given depth. In order to apply Theorem 3.4 we show that also for 2-identification the uniform distribution is optimal for saturated block codes. With these results we obtain an expression as a lower bound which still depends on (1-)identification. However, the (1-)identification running time appears negatively signed so that we cannot immediately apply its lower bound. This lower bound is the identification entropy $H_{I,q}$ established in [3]. During this thesis we refer to $H_{\mathrm{ID}}^{1,q} = H_{I,q}$ since, as we will see, identification entropy is a special case of the q-ary identification entropy of degree L.

In the beginning of Subsection 4.2 we show that if the underlying probability distribution consists only of q-powers, the previously established lower bound can be attained. This ensures us to define the q-ary identification entropy of second degree by

$$H_{\mathrm{ID}}^{2,q}(P) = 2\frac{q}{q-1}\left(1 - \sum_{u\in\mathcal{U}} p_u^2\right) - \frac{q^2}{q^2-1}\left(1 - \sum_{u\in\mathcal{U}} p_u^3\right).$$

This function obeys some important properties, which appear as desiderata for entropy functions in [1]. It is symmetric, normalized, decisive and expansible. Further, it is lower bounded by the probability distribution where all the probability is concentrated in one point and upper bounded by the uniform distribution. Finally, we establish a grouping behavior, which is a generalization of the grouping behavior of the identification entropy function $H_{\mathrm{ID}}^{1,q}$. With these properties we finally prove that $H_{\mathrm{ID}}^{2,q}$ is indeed a lower bound for the symmetric running time for q-ary 2-identification. Moreover, we show that this bound can be attained if and only if P consists only of q-powers. As a final result of this subsection we show that balanced Huffman codes are asymptotically optimal for 2-identification.

In the final subsection we provide an upper bound for the worst-case running time by the same code construction which we used in Subsection 2.2.

In the following Section 5 we turn to L-identification for general distributions and define the q-ary identification entropy of degree L by

$$H_{\mathrm{ID}}^{L,q}(P) = -\sum_{l=1}^{L}(-1)^l \binom{L}{l} \frac{q^l}{q^l-1}\left(1-\sum_{u\in\mathcal{U}} p_u^{l+1}\right).$$

We show that also this entropy function is symmetric, normalized, decisive and expansible. It further obeys a grouping behavior, which is a generalized version of the previous grouping behavior for $L = 1, 2$. Unfortunately, we were not able to prove a lower and upper bound. There exist counterexamples for which uniform distribution is not an upper bound. These counterexamples only occur if $N < q$, i.e. the size of the output space is strictly less than the alphabet size. However, in order to show that $H_{\mathrm{ID}}^{L,q}$ is a lower bound for L-identification we only need the bounds for the case $N = q$. We prove this relation under the assumption that in this case uniform distribution is indeed an upper bound. If, additionally, we assume that it is the only distribution which attains this upper bound we can show that there exists a code $\mathcal{C}$ with $H_{\mathrm{ID}}^{L,q}(P) = \mathcal{L}_{\mathcal{C}}^{L,q}(P,P)$ if and only if P consists only of q-powers.

In Section 6 we turn to another type of identification namely L-identification for sets. We begin by defining L-identification for sets and point out the differences to L-identification (for vectors). After that we show that if we consider the uniform distribution and balanced Huffman codes, the symmetric running time of L-identification for sets asymptotically equals the symmetric running time of L-identification.

In the final Section 7 we state some open problems which arose during the analysis of L-identification.

1 Definitions and Notation

In this section we provide definitions and notations, which are the base for all further calculations. The first subsection is a short overview of source coding. We further introduce code trees, which are useful for visualizing behaviors of a given code. In the second subsection we explain the task of an L-identification code and define the performance behaviors in which we are interested in this thesis.

We begin with some set-theoretical notation. The set of the natural numbers 1 to n is denoted by $[n]$ and the set of all natural numbers from $m+1$ up to n is denoted by $[m+1, n]$. However, $[0, 1]$ still denotes the closed real interval from 0 to 1. Let $\mathcal{S}$ be any finite set. Then $2^{\mathcal{S}}$ denotes the power set of $\mathcal{S}$, $\binom{\mathcal{S}}{k}$ denotes the set of all k-element subsets of $\mathcal{S}$ and $\mathcal{S}^* = \bigcup_{d=0}^{\infty} \mathcal{S}^d$. Further, let P be a probability distribution on $\mathcal{S}$. Then, $supp(P) = \{s \in \mathcal{S} \ : \ P(\{s\}) \neq 0\}$ denotes the *support* of P.

We often have to deal with functions whose arguments are probability distributions on some given finite set. Therefore we formally define a domain for these functions. Following [1] (pp. 26) we define

$$\Delta_n = \{(p_1, ..., p_n) \in [0,1]^n \ : \ 0 \leq \sum_{i=1}^{n} p_i \leq 1\}$$

to be the set of all (perhaps incomplete) probability distributions on $[n]$ and

$$\Gamma_n = \{(p_1, ..., p_n) \in [0,1]^n \ : \ \sum_{i=1}^{n} p_i = 1\}$$

to be the set of all complete probability distributions on $[n]$. If we want to exclude zero probabilities, we write for $n \geq 2$

$$\mathring{\Delta}_n = \{(p_1, ..., p_n) \in (0,1)^n \ : \ 0 < \sum_{i=1}^{n} p_i \leq 1\}$$

and

$$\mathring{\Gamma}_n = \{(p_1, ..., p_n) \in (0,1)^n \ : \ \sum_{i=1}^{n} p_i = 1\}.$$

It follows immediately from the above definitions that

$$\Gamma_n = \{(p_1, ..., p_n) \in (0,1)^n \ : \ (p_1, ..., p_{n-1}) \in \Delta_{n-1} \text{ and } p_n = 1 - \sum_{i=1}^{n-1} p_i\}. \quad (1.1)$$

This means that Γ_n is a $(n-1)$-dimensional hyperplane in the n-dimensional real space. Hence, if we analyze a function $f : \Gamma_n \to \mathbb{R}$ by differentiation, we only have to consider $n-1$ partial derivatives

$$\frac{\delta}{\delta x_j} \tilde{f}\,(p_1, ..., p_{n-1})\,,$$

with $j \in [n-1]$ and where $\tilde{f}(p_1, ..., p_{n-1}) = f(p_1, ..., p_{n-1}, 1 - \sum_{i=1}^{n-1} p_i)$.

For a mapping $f : \Gamma_n \to \mathbb{R}$ we will write $f(P) = f(p_1, ..., p_N)$. Thus, omitting the additional brackets on the right hand side. For a function $g : \Gamma_n^2 \to \mathbb{R}$, however, we retain the brackets and write $g(P, R) = g((p_1, ..., p_N), (r_1, ..., r_N))$.

1.1 Source Coding and Code Trees

A *discrete source* is a probability space $(\mathcal{U}, 2^{\mathcal{U}}, P)$, where $\mathcal{U}$ is a finite set, called the *output space*. W.l.o.g. we assume that $\mathcal{U} = [N]$ for some $N \in \mathbb{N}$. Further, P is a probability distribution on $\mathcal{U}$ with $p_u = P(\{u\})$. It is called the *output probability distribution*. Often, the indication of $2^{\mathcal{U}}$ is omitted and we will follow this standard and call $(\mathcal{U}, P)$ a discrete source with output space $\mathcal{U} = [N]$ and output probability P. We further introduce the *output random variable* $U = id_{\mathcal{U}}$. It follows that $Prob(U = u) = p_u$.

A discrete *memoryless* source $(\mathcal{U}^n, P^n)$ is characterized by $P_{u^n} = P^n(\{u^n\}) = \prod_{i=1}^n p_{u_i}$ for all $u^n = (u_1, u_2, .., u_n)$. $U^n = id_{\mathcal{U}^n}$ is the output random variable for this discrete memoryless source.

For the *alphabet* $\mathcal{Q} = \{0, 1, ..., q-1\}$ a mapping $\mathcal{C} : \mathcal{U} \to \mathcal{Q}^*$ is a called *q-ary code* on $\mathcal{U}$ and $\mathcal{C}(u) = c_u = c_{u,1}c_{u,2}...c_{u,\|c_u\|}$ is the q-ary codeword of $u \in \mathcal{U}$. The individual $c_{u,i} \in \mathcal{Q}$ are called q-bits. We also write shortly that $\mathcal{C} = \{c_1, ..., c_N\}$. Further, for $u \in \mathcal{U}$ and $k \in [\|c_u\| - 1]$ we define $c_u^k = c_{u,1}...c_{u,k}$ to be the *prefix* of length k of the codeword c_u. In addition we set $c_u^0 = e$, where e is the empty codeword.

A code is called a *prefix code* if no codeword is prefix of another. Formally, for each $c \in \mathcal{C}$ let

$$D(c) = \bigcup_{k=0}^{\|c\|-1} c^k. \quad (1.2)$$

Then, $\mathcal{C}$ is a prefix code if and only if it holds for all $c, c' \in \mathcal{C}$ that $c \notin D(c')$. For more information on prefix codes we refer to [17]. Hereafter, unless otherwise specified, by a code we shall always understand a prefix code. We also define for some code $\mathcal{C}$ the set of all prefixes of its codewords by

$$D(\mathcal{C}) = \bigcup_{c \in \mathcal{C}} D(c). \tag{1.3}$$

A *block code* is a code where all codewords have the same length. We further use $\mathcal{C}_{q^n}$ to denote the q-ary block code of size q^n. It is a special block code and called *saturated*.

It is often useful to visualize a code by its *code tree*. Therefore consider a q-ary tree, where all branches with the same branching point are labeled with elements of $\mathcal{Q}$. Such a tree is a code tree $T_{\mathcal{C}}$ of a code $\mathcal{C}$ if there exists a bijective mapping ϕ from the set $\bar{\mathcal{N}}(T_{\mathcal{C}})$ of leaves of $T_{\mathcal{C}}$ onto $\mathcal{C}$ such that $\phi(x)$ equals the labeled path from the root of $T_{\mathcal{C}}$ to leaf x for all $x \in \bar{\mathcal{N}}(T_{\mathcal{C}})$. Figure 1.1 shows an example of a code and its corresponding code tree.

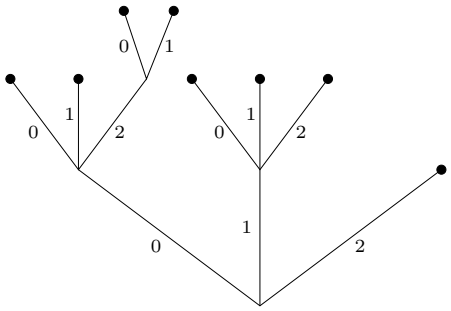

Figure 1.1: The ternary code tree of $\mathcal{C} = \{00, 01, 020, 021, 10, 11, 12, 2\}$.

We have already used the expression $\bar{\mathcal{N}}(T)$ for the set of leaves or external nodes. In addition, we use $\mathring{\mathcal{N}}(T)$ for the set of branching points or inner nodes of a tree T and $\mathcal{N}(T) = \bar{\mathcal{N}}(T) \cup \mathring{\mathcal{N}}(T)$. The bijective mapping ϕ from before can be extended to $\mathcal{N}(T)$ by mapping every inner node $x \in \mathring{\mathcal{N}}(T)$ to the element in $D(\mathcal{C})$ which corresponds to the labeled path from the root of T to x. Because of this direct connection we do not distinguish between a code and its code tree. We will use $\mathcal{C}$ and $\bar{\mathcal{N}}(T_{\mathcal{C}})$ equivalently and the same we do for $D(\mathcal{C})$ and $\mathring{\mathcal{N}}(T_{\mathcal{C}})$.[1] That is, we equivalently use x and $\phi(x)$. For example, $\|x\| = \|\phi(x)\|$. Further, T_x (or $T_{\phi(x)}$) denotes the subtree of T with root in x for some node $x \in \mathcal{N}(T)$.

[1] This can only be done because we consider prefix codes.

If $\|x\| = 0$, then $T_x = T_e = T$, and if $x \in \bar{\mathcal{N}}(T)$, then $T_x = x$.

Let $\mathcal{C}$ be a source code for the source $(\mathcal{U}, P)$. The *concatenated code* $\mathcal{C}^n$ for the source $(\mathcal{U}^n, P^n)$ is defined as follows. The codeword for each output $u^n = (u_1, ..., u_n)$ is the concatenation of the individual codewords of the u_i's. That is

$$c_{u^n} = c_{u_1} ... c_{u_n}.$$

If we consider a concatenated code $\mathcal{C}^n$, then $\mathcal{C}$ is called the *basic code*. $\mathcal{C}^n$ can also be obtained by a stepwise construction. Therefore consider the code tree $T_{\mathcal{C}}$. For each *concatenation step* $1 \leq t \leq n-1$ the new code tree $T_{\mathcal{C}^{t+1}}$ is obtained by replacing each of the leaves of $T_{\mathcal{C}^t}$ with a copy of $T_{\mathcal{C}}$. Figure 1.2 shows the first concatenation step of a binary code by means of its code tree. Every node of the concatenated tree where two basic trees are connected is called a *concatenation point*.

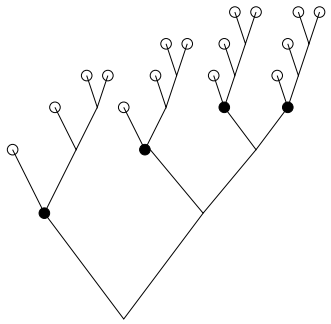

Figure 1.2: The concatenated tree $T_{\mathcal{C}^2}$ corresponding to the binary code $\mathcal{C} = \{0, 10, 110, 111\}$.

1.2 L-Identification

Consider the discrete memoryless source $(\mathcal{U}^L, P^L)$ together with a source code $\mathcal{C}$ on $\mathcal{U}$. Additionally and in contrast to classical source coding we also introduce the so-called user space $\mathcal{V}$, with $|\mathcal{V}| = |\mathcal{U}|$, together with the *user random variable* $V = id_{\mathcal{V}}$. Let $f : \mathcal{V} \to \mathcal{U}$ be a bijective mapping. We encode the *users* v with the same code $\mathcal{C}$ as before. That is, we set $c_v = c_{f(v)}$. W.l.o.g. we assume from now on that $\mathcal{V} = \mathcal{U}$ and $f = id_{\mathcal{U}}$.

The task of L-identification is to decide for every user $v \in \mathcal{U}$ and every output $u^L = (u_1, .., u_L) \in \mathcal{U}^L$ whether or not there exists at least one $l \in [L]$ such that

$v = u_l$. To achieve this goal we compare step by step the first, second, third etc. q-bit of c_v with the corresponding q-bits of $c_{u_1}, ..., c_{u_L}$. After each step i all u_l with $c_{u_l,i} \neq c_{v,i}$ are eliminated from the set of possible candidates. We continue with step $i+1$ comparing only those u_l which still are candidates. If at some point during this procedure the last possible candidate is eliminated, the L-identification process stops and returns "No, v is not contained in u^L.". On the other hand, if there are still candidates after the comparison of the last q-bit of c_v, the L-identification process also halts but returns "Yes, v is contained in u^L at position(s) ...". The number of steps until the process halts is called the *L-identification running time* for $(u^L, v) \in \mathcal{U}^L \times \mathcal{U}$.

The algorithm `LID` presented in the appendix in Table 8.1 accomplishes L-identification. As its input serve the codewords $c_{u_1}, ..., c_{u_L}$ and c_v and it returns the triple $(A, s, \mathcal{S})$. Here A is a boolean variable which is "TRUE" if v is contained in u^L and "FALSE" if not. The second component s equals the number of steps until the algorithm halted and the third component returns the set of positions of the output vector u^L which coincide with the user v. This means that if there exist one or more components of u^L which coincide with v, we also know their exact number and positions. This is not a requirement to L-identification but an extra feature. It follows from the fact that up to the last comparison of q-bits still all possible candidates may not coincide with v.

In Section 6 we turn to L-identification for sets and there this feature is not attained since we know that all still possible candidates are pairwise distinct. This means that in some cases L-identification for sets can be faster than L-identification (for vectors). In this case, however, we do not know where the particular user occurred. We explain what we mean by L-identification for sets and point out the differences in greater detail in Section 6.

Formally, we define the L-identification running time for given u^L, v and q-ary code $\mathcal{C}$ by

$$\mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v) = \mathtt{LID}_2(c_{u_1}, ..., c_{u_L}, c_v), \tag{1.4}$$

where $\mathtt{LID}_2(c_{u_1}, ..., c_{u_L}, c_v)$ is the second component of the triple returned by the algorithm `LID`.

The goal of this thesis is to analyze the expected length of the L-identification running time, also called the *average running time*, for a given user $v \in \mathcal{U}$

$$\mathcal{L}_{\mathcal{C}}^{L,q}(P, v) = \sum_{u^L \in \mathcal{U}^L} P_{u^L} \mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v). \tag{1.5}$$

This can be done in different ways. The first is the worst-case scenario where we are interested in the *worst-case average running time*, which we shortly call

the *worst-case running time*,

$$\mathcal{L}_{\mathcal{C}}^{L,q}(P) = \max_{v \in \mathcal{U}} \mathcal{L}_{\mathcal{C}}^{L,q}(P, v). \tag{1.6}$$

We want to find codes which are as close as possible to the *optimal worst-case running time*

$$\mathcal{L}^{L,q}(P) = \min_{\mathcal{C}} \mathcal{L}_{\mathcal{C}}^{L,q}(P). \tag{1.7}$$

In Subsections 2.2 and 4.3 we provide upper bounds for $\mathcal{L}^{1,2}(P)$ and $\mathcal{L}^{2,2}(P)$.

Let us assume that also user v is chosen at random according to a probability distribution Q on $\mathcal{U}$. We are now interested in the *expected average running time* or shortly the *expected running time*

$$\mathcal{L}_{\mathcal{C}}^{L,q}(P, Q) = \sum_{v \in \mathcal{U}} Q(\{v\}) \mathcal{L}_{\mathcal{C}}^{L,q}(P, v) \tag{1.8}$$

and in particular in the *optimal expected running time*

$$\mathcal{L}^{L,q}(P, Q) = \min_{\mathcal{C}} \mathcal{L}_{\mathcal{C}}^{L,q}(P, Q). \tag{1.9}$$

In this thesis we focus on the special case where $Q = P$ so that Equations (1.8) and (1.9) become

$$\mathcal{L}_{\mathcal{C}}^{L,q}(P, P) = \sum_{v \in \mathcal{U}} p_v \mathcal{L}_{\mathcal{C}}^{L,q}(P, v) = \sum_{(u^L, v) \in \mathcal{U}^{L+1}} P_{u^L} p_v \mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v) \tag{1.10}$$

and

$$\mathcal{L}^{L,q}(P, P) = \min_{\mathcal{C}} \mathcal{L}_{\mathcal{C}}^{L,q}(P, P). \tag{1.11}$$

We call $\mathcal{L}_{\mathcal{C}}^{L,q}(P, P)$ the *symmetric running time* for a given code $\mathcal{C}$ and $\mathcal{L}^{L,q}(P, P)$ the *optimal symmetric running time*. In Section 4 we derive an entropy function for 2-identification. This function provides a lower bound for $\mathcal{L}^{2,q}(P, P)$. In Section 5 we discuss an extension of this approach to the case of L-identification for general L. It is clear from the above definitions that

$$\mathcal{L}^{L,q}(P, P) \leq \mathcal{L}^{L,q}(P) \tag{1.12}$$

so that the bounds we derive in Section 5 and Subsections 2.2, 4.2 and 4.3 are lower (resp. upper) bounds for both values.

All the above values also depend on $N = |\mathcal{U}|$. We do not state this fact explicitly since it is contained in both P and $\mathcal{C}$.

2 Two new results for (1-)Identification

In this section we state two new results for (1-)identification. The first result is about (1-)identification for block codes. In [5] it is proven that the q-ary identification entropy $H_{\mathrm{ID}}^{1,q}(P)$ is a lower bound for $\mathcal{L}_{\mathcal{C}}^{1,q}(P,P)$. A key step in this proof is to show that if $\mathcal{C}$ is a saturated block code, the running time of identification is minimized by the uniform distribution. This result is provided in Subsection 2.1. Although this may seem obvious the proof is not trivial. Moreover, we will see in Section 5 that at least for $L \geq 4$ the uniform distribution is not always optimal for L-identification on block codes.

The second result is about upper bounds for the worst-case running time. In Section 4 of [4] the authors proved in Theorem 3 that $\mathcal{L}^{1,2}(P) < 3$ by an inductive code construction. We discovered that with a small alteration of their construction this upper bound can be strengthened.

2.1 (1-)Identification for Block Codes

In order to show that the uniform distribution is optimal for (1-)identification on block codes we modify any given probability distribution step by step until we reach the uniform distribution without increasing $\mathcal{L}_{\mathcal{C}_{q^n}}^{1,q}(P,P)$. It turns out that not only the uniform distribution is optimal. In fact, all distributions $P = (p_1, ..., p_{q^n})$ are optimal for which we are able to partition $\mathcal{U} = [q^n]$ into sets $\mathcal{U}_1, ..., \mathcal{U}_{q^{n-1}}$, all of cardinality q, such that $\sum_{u \in \mathcal{U}_i} p_u = 1/q^{n-1}$ for all $i \in [q^{n-1}]$. This is due to the fact that the running time regarding v is the same for all u whose codewords c_u coincide with c_v in all but the last q-bit. The individual steps of modification and their monotone decreasing property are content of

Lemma 2.1 *Let $n \in \mathbb{N}$, $q \in \mathbb{N}_{\geq 2}$, $k \in \{0, ..., n-1\}$ and $t \in \{0, ..., q^{n-k-1} - 1\}$. Further, let $P = (p_1, ..., p_{q^n})$ and $\tilde{P} = (\tilde{p}_1, ..., \tilde{p}_{q^n})$ be probability distributions on $[q^n]$ with*

$$P = (p_1, ..., p_{tq^{k+1}}, \underbrace{r_1, ..., r_1}_{q^k}, \underbrace{r_2, ..., r_2}_{q^k}, ..., \underbrace{r_q, ..., r_q}_{q^k}, p_{(t+1)q^{k+1}+1}, ..., p_{q^n})$$

and

$$\tilde{P} = (p_1, ..., p_{tq^{k+1}}, \underbrace{\frac{1}{q}\sum_{i=1}^{q} r_i, ..., \frac{1}{q}\sum_{i=1}^{q} r_i}_{q^{k+1}}, p_{(t+1)q^{k+1}+1}, ..., p_{q^n}).$$

Then it holds

$$\mathcal{L}^{1,q}_{\mathcal{C}_{q^n}}(P, P) - \mathcal{L}^{1,q}_{\mathcal{C}_{q^n}}(\tilde{P}, \tilde{P}) = \frac{q^k(q^k-1)}{2(q-1)} \sum_{i,j=1}^{q} (r_i - r_j)^2 \geq 0.$$

The inequality holds with equality if and only if either $k = 0$ or $r_i = r_j$ for all $i, j \in [q]$.

Proof:

W.l.o.g. we assume that $t = 0$, such that

$$P = (p_1, ..., p_{q^n}) = (r_1, ..., r_1, r_2, ..., r_2, ..., r_q, ..., r_q, p_{q^{k+1}+1}, ..., p_{q^n})$$

and

$$\tilde{P} = (\tilde{p}_1, ..., \tilde{p}_{q^n}) = (\frac{1}{q}\sum_{i=1}^{q} r_i, ..., \frac{1}{q}\sum_{i=1}^{q} r_i, p_{q^{k+1}+1}, ..., p_{q^n})$$

Also, we use for simplicity the abbreviations $L_{u,v} = \mathcal{L}^{1,q}_{\mathcal{C}_{q^n}}(u, v)$ and $\alpha_{u,v} = (p_u p_v - \tilde{p}_u \tilde{p}_v) L_{u,v}$. It is clear that $L_{u,v} = L_{v,u}$ and hence $\alpha_{u,v} = \alpha_{v,u}$. Also, $\alpha_{u,v} = 0$ for all $u, v \in [q^{k+1} + 1, q^n]$. This yields

$$\mathcal{L}^{1,q}_{\mathcal{C}_{q^n}}(P, P) - \mathcal{L}^{1,q}_{\mathcal{C}_{q^n}}(\tilde{P}, \tilde{P}) = \sum_{u,v=1}^{q^n} \alpha_{u,v} = \sum_{u,v=1}^{q^{k+1}} \alpha_{u,v} + 2\sum_{u=1}^{q^{k+1}} \sum_{v=q^{k+1}+1}^{q^n} \alpha_{u,v}.$$

It further holds for $u \in [q^{k+1}]$ and $v \in [q^{k+1} + 1, q^n]$ that

i) $p_v = \tilde{p}_v$,

ii) $L_{u,v} = L_{1,v}$, which we denote by L_v,

iii) $\tilde{p}_u = \frac{1}{q}\sum_{i=1}^{q} r_i$ and

iv) $\sum_{u=1}^{q^{k+1}} p_u = q^k \sum_{i=1}^{q} r_i$.

From iii) and iv) it follows that

$$\sum_{u=1}^{q^{k+1}} \sum_{v=q^{k+1}+1}^{q^n} \alpha_{u,v} = \sum_{v=q^{k+1}+1}^{q^n} p_v L_v \sum_{u=1}^{q^{k+1}} (p_u - \tilde{p}_u) = 0$$

and hence

$$\begin{aligned} & \mathcal{L}^{1,q}_{\mathcal{C}_{q^n}}(P,P) - \mathcal{L}^{1,q}_{\mathcal{C}_{q^n}}(\tilde{P},\tilde{P}) \\ = & \sum_{u,v=1}^{q^{k+1}} \left[p_u p_v - \tfrac{1}{q^2} \left(\sum_{i=1}^{q} r_i \right)^2 \right] L_{u,v} \\ = & \sum_{j,m=1}^{q} \left[r_j r_m - \tfrac{1}{q^2} \left(\sum_{i=1}^{q} r_i \right)^2 \right] \sum_{u=(j-1)q^k+1}^{jq^k} \sum_{v=(m-1)q^k+1}^{mq^k} L_{u,v}. \end{aligned}$$

Here, the first equality follows from iii) and the definition of $\tilde{P}$. The second equality is due to the definition of P.

We now take a look at $L_{u,v}$ and see that for $u \in [(j-1)q^k + 1, jq^k]$ and $v \in [(m-1)q^k + 1, mq^k]$ we have

$$L_{u,v} = \begin{cases} n-k & \text{if } j \neq m \\ n - k + \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u,v) & \text{if } j = m. \end{cases}$$

With this observation we get

$$\begin{aligned} & \mathcal{L}^{1,q}_{\mathcal{C}_{q^n}}(P,P) - \mathcal{L}^{1,q}_{\mathcal{C}_{q^n}}(\tilde{P},\tilde{P}) \\ = & (n-k)q^{2k} \sum_{j,m=1}^{q} \left[r_j r_m - \tfrac{1}{q^2} \left(\sum_{i=1}^{q} r_i \right)^2 \right] \\ & + \sum_{j=1}^{q} \left[r_j^2 - \tfrac{1}{q^2} \left(\sum_{i=1}^{q} r_i \right)^2 \right] \sum_{u,v=1}^{q^k} \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u,v) \qquad (2.1) \\ = & \sum_{j=1}^{q} \left[r_j^2 - \tfrac{1}{q^2} \left(\sum_{i=1}^{q} r_i \right)^2 \right] q^k \left[(q-1)q^k \sum_{l=1}^{k} l q^{-l} + k \right] \\ = & \tfrac{q}{q-1} q^k (q^k - 1) \sum_{j=1}^{q} \left[r_j^2 - \tfrac{1}{q^2} \left(\sum_{i=1}^{q} r_i \right)^2 \right]. \end{aligned}$$

The first equality follows from the additional fact that $\sum_{u,v=(j-1)q^k+1}^{jq^k} L_{u,v}$ is invariant in the choice of $j \in [q]$. The partial sum behavior of the geometric series yields the third equality. To understand the second equality we see that

$$\sum_{j,m=1}^{q} \left[r_j r_m - \frac{1}{q^2} \left(\sum_{i=1}^{q} r_i \right)^2 \right] = \sum_{j,m=1}^{q} r_j r_m - \left(\sum_{i=1}^{q} r_i \right)^2 = 0.$$

In addition, we have that

$$\sum_{u,v=1}^{q^k} \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u,v) = \sum_{v=1}^{q^k} \sum_{l=1}^{k} l | \{ u \in [q^k] \ : \ \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u,v) = l \} |.$$

For $l = 1, ..., k-1$ the codeword of each element in the above sets has to coincide with c_v in the first $l-1$ q-bits. Those are q^{k-l+1} many. Furthermore, each one of those codewords has to differ from c_v in the l-th q-bit. These are $q-1$ out of q. We end up with $(q-1)q^{k-l}$ elements. If $l = k$, also v itself is contained in the corresponding set. As one can see, this is invariant of the choice of $v \in [q^k]$. It follows that

$$\begin{aligned} \sum_{v=1}^{q^k} \sum_{l=1}^{k} l | \{ u \in [q^k] \ : \ \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u,v) = l \} | &= q^k \left[\sum_{l=1}^{k-1} l(q-1)q^{k-l} + kq \right] \\ &= q^k \left[(q-1)q^k \sum_{l=1}^{k} l q^{-l} + k \right]. \end{aligned}$$

This proves the second equality of Equation (2.1). Finally, since

$$\sum_{j=1}^{q} \left[r_j^2 - \frac{1}{q^2} \left(\sum_{i=1}^{q} r_i \right)^2 \right] = \frac{1}{2q} \sum_{i,j=1}^{q} (r_i - r_j)^2,$$

we obtain the expression to be proven.

□

Lemma 2.1 provides a way to come step by step from any given distribution $P = (p_1, ..., p_{q^n})$ to the uniform distribution without increasing the symmetric 2-identification running time on q-ary block codes. In the first step ($t = 0$) of the first round ($k = 0$) we level out the probabilities $p_1, ..., p_q$. In the second step ($t = 1$, $k = 0$) we level out $p_{q+1}, ..., p_{2q}$ and so on until in the last step ($t = q^{n-1} - 1$) of the first round the remaining probabilities p_{q^n-q+1} up to p_{q^n} are leveled out. We have not changed the symmetric 2-identification running time, and we have constructed a probability distribution which enables us to go

on with Lemma 2.1. This is due to the fact that the first q , the second q up to the last q probabilities are now identical. In round 2 ($k = 1$) we begin to level out the first q^2 probabilities, then the second q^2 probabilities up to the last q^2. During these actions Lemma 2.1 ensures us that the symmetric 2-identification running time does not increase. Again we end up with a distribution which allows us to apply Lemma 2.1 also in the third round $k = 2$ and so on. Finally, in the last round $k = n-1$ we level out the first q^{n-1} identical probabilities and the second and last q^{n-1} identical probabilities and end up with the uniform distribution. We have proven the following

Corollary 2.2 *Let $n \in \mathbb{N}$ and $q \in \mathbb{N}_{\geq 2}$. Further, let $\mathcal{C} = \mathcal{C}_{q^n}$ and $T = T_{\mathcal{C}}$. Then, for all probability distributions P on $[q^n]$ it holds that*

$$\mathcal{L}_{\mathcal{C}}^{1,q}(P,P) \geq \mathcal{L}_{\mathcal{C}}^{1,q}\left((\frac{1}{q^n},...,\frac{1}{q^n}),(\frac{1}{q^n},...,\frac{1}{q^n})\right),$$

with equality if and only if $P(T_x) = q^{-\|x\|}$ for all inner nodes $x \in \mathring{\mathcal{N}}(T)$.

2.2 An Improved Upper Bound for Binary Codes

In Section 4 of [4] the authors proved in Theorem 3 that $\mathcal{L}^{1,2}(P) < 3$ by an inductive code construction. They assumed that w.l.o.g. $p_1 \geq p_2 \geq ... \geq p_N$. In the first step $\mathcal{U}$ is partitioned into $\mathcal{U}_0 = [t]$ and $\mathcal{U}_1 = [t+1, N]$ such that $\sum_{i=1}^{t} p_i$ is as close as possible to $1/2$. Then, they inductively construct code on $\mathcal{U}_0$ and $\mathcal{U}_1$. Finally, that they prefixed the codewords for all elements in $\mathcal{U}_0$ (resp. $\mathcal{U}_1$) by **0** (resp. **1**).

The proof of this theorem contains some cases differentiations. The worst of these cases is that $\sum_{i=1}^{t} p_i < \frac{1}{2}$ and the user $v_{\max}$ which maximizes $\mathcal{L}_{\mathcal{C}}^{1,2}(P,v)$ is in $\mathcal{U}_1$. [1] In this case we may take up to a certain number additional outputs from $\mathcal{U}_1$ and put them into $\mathcal{U}_0$ in order to speed up the identification process. To do so we define

$$\mathcal{U}_{\max} = \{u \in \mathcal{U} \ : \ c_{u,1} = c_{v_{\max},1}\} \tag{2.2}$$

and

$$p_{\max} = \sum_{u \in \mathcal{U}_{\max}} p_u. \tag{2.3}$$

Further, $P_{\max}$ is a probability distribution on $\mathcal{U}_{\max}$ defined by

$$P_{\max,u} = \frac{p_u}{p_{\max}} \tag{2.4}$$

[1] $v_{\max}$ may not be unique, but if there are more than one, it does not matter which of these we choose.

for all $u \in \mathcal{U}_{\max}$ and $\mathcal{C}_{\max}$ is the code on $\mathcal{U}_{\max}$ which we obtain by deleting the leading bit of all c_u's. With these definitions we get that

$$\begin{aligned}
\mathcal{L}_{\mathcal{C}}^{L,q}(P) &= \sum_{u^L \in \mathcal{U}^L} p_{u_1} ... p_{u_L} \mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v_{\max}) \\
&= 1 + \sum_{l=1}^{L} \binom{L}{l} (1 - p_{\max})^{L-l} \sum_{u^l \in \mathcal{U}_{\max}^l} p_{u_1} ... p_{u_l} \mathcal{L}_{\mathcal{C}}^{l,q}(u^l, v_{\max}) \\
&= 1 + \sum_{l=1}^{L} \binom{L}{l} (1 - p_{\max})^{L-l} p_{\max}^l \mathcal{L}_{\mathcal{C}}^{l,q}(P_{\max}, v_{\max}) \\
&\leq 1 + \sum_{l=1}^{L} \binom{L}{l} (1 - p_{\max})^{L-l} p_{\max}^l \mathcal{L}_{\mathcal{C}_{\max}}^{l,q}(P_{\max}).
\end{aligned} \tag{2.5}$$

This simplifies for $L = 1$ and $q = 2$ to

$$\mathcal{L}_{\mathcal{C}}^{1,2}(P) \leq 1 + p_{\max} \mathcal{L}_{\mathcal{C}_{\max}}^{1,2}(P_{\max}). \tag{2.6}$$

This equation provides the induction step for the proof of

Theorem 2.3 *It holds for all probability distributions P on $\mathcal{U}$ that the worst-case running time for binary (1-)identification can be upper bounded by*

$$\mathcal{L}^{1,2}(P) < \frac{5}{2}.$$

Proof:

W.l.o.g. we assume that $p_1 \geq p_2 \geq ... \geq p_N$. For the induction bases $N = 1, 2$ we have that $\mathcal{L}^{1,2}(P) = 1 < 5/2$ for all P. Now let $N > 2$.

Case 1: $\mathbf{p_1 \geq \frac{1}{2}}$

In this case we assign $c_1 = \mathbf{0}$ and $\mathcal{U}_1 = \{2, ..., N\}$. Inductively we construct a code $\mathcal{C}' = \{c'_u \ : \ u = 2, ..., N\}$ on $\mathcal{U}_1$ and we extend this code to a code on $\mathcal{U}$ by setting $c_u = \mathbf{1}c'_u$ for $u \in \mathcal{U}_1$.

It is clear that $v_{\max} \neq 1$ because in this case $\mathcal{L}^{1,2}(P)$ would equal 1. This is a contradiction since $N > 2$ and thereby we have more than one output whose codeword begins with **1** and each of these outputs results in a running time strictly greater than 1.

Thus, the maximum is assumed on the "right" side. This yields $p_{\max} \leq 1/2$. Further, by Equation (2.6) and the induction hypothesis we have that

$$\mathcal{L}_{\mathcal{C}}^{1,2}(P) < 1 + \frac{1}{2} \cdot \frac{5}{2} = \frac{9}{4} < \frac{5}{2}.$$

Case 2: $\mathbf{p_1 < \frac{1}{2}}$

In this case we choose t such that $|1/2 - \sum_{u=1}^{t} p_u|$ is minimized. Now we distinguish again between two subcases.

Case 2.1: $\mathbf{t = 1}$

In this case we set $\mathcal{U}_0 = \{1, 2\}$ and $\mathcal{U}_1 = \{3, ..., N\}$. Again by we inductively construct $\mathcal{C}' = \{c'_u \ : \ u = 3, ..., N\}$. And we obtain $\mathcal{C}$ by setting $c_1 = \mathbf{00}$, $c_2 = \mathbf{01}$ and $c_u = \mathbf{1}c'_u$ for $u = 3, ..., N$.

If $v_{\max} \in \mathcal{U}_0$, we have that $p_{\max} = p_1 + p_2$ and $\mathcal{C}_{\max} = \{\mathbf{0}, \mathbf{1}\}$. Again by equation (2.6) we obtain

$$\mathcal{L}_{\mathcal{C}}^{1,2}(P) \leq 1 + (p_1 + p_2)\mathcal{L}_{\mathcal{C}_{\max}}^{1,2}(P_{\max}) \leq 2 < \frac{5}{2}.$$

Otherwise it follows from the definition of t that $p_1 + p_2 > 1/2$. By this we get $p_{\max} < 1/2$ and $\mathcal{C}_{\max} = \mathcal{C}_1$. By induction and Equation (2.6) this yields

$$\mathcal{L}_{\mathcal{C}}^{1,2}(P) < 1 + \frac{1}{2} \cdot \frac{5}{2} = \frac{9}{4} < \frac{5}{2}.$$

Case 2.2: $\mathbf{t \geq 2}$

We now set $\mathcal{U}_0 = \{1, ..., t\}$ and $\mathcal{U}_1 = \{t + 1, ..., N\}$ and construct inductively codes $\mathcal{C}' = \{c'_u \ : \ u = 1, ..., t\}$ and $\mathcal{C}'' = \{c''_u \ : \ u = t + 1, ..., N\}$. We obtain a code $\mathcal{C}$ on $\mathcal{U}$ by setting

$$c_u = \begin{cases} \mathbf{0}c'_u & \text{for } u = 1, ..., t \\ \\ \mathbf{1}c''_u & \text{for } u = t + 1, ..., N. \end{cases}$$

Case 2.2.1: $\mathbf{v_{\max} \in \mathcal{U}_0}$

It follows that $p_{\max} = \sum_{u=1}^{t} p_u$. If $\sum_{u=1}^{t} p_u \leq 1/2$, we get again by induction and Equation (2.6) that

$$\mathcal{L}_{\mathcal{C}}^{1,2}(P) < 1 + \frac{1}{2} \cdot \frac{5}{2} = \frac{9}{4} < \frac{5}{2}.$$

In the case that $\sum_{u=1}^{t} p_u > 1/2$ we have by the definition of t that

$$\sum_{u=1}^{t} p_u - \frac{1}{2} \leq \frac{1}{2} - \sum_{u=1}^{t-1} p_u.$$

It follows $\sum_{u=1}^{t} p_u \leq (p_t+1)/2$. Additionally, we have $p_{t-1} < 1/(2(t-1))$ because otherwise $\sum_{u=1}^{t-1} p_u \geq 1/2$. This would be a contradiction to the definition of t. This together implies

$$p_{\max} = \sum_{u=1}^{t} p_u < \frac{1 + 2(t-1)}{4(t-1)}. \tag{2.7}$$

If $t = 2$, we obtain for the same reasons as in Case 2.1 that

$$\mathcal{L}_{\mathcal{C}}^{1,2}(P) < \frac{5}{2}.$$

If $t = 3$, we get that $\mathcal{C}_{\max} = \mathcal{C}' = \{c'_1, c'_2, c'_3\}$, with $c'_1 = \mathbf{0}$, $c'_2 = \mathbf{10}$ and $c'_3 = \mathbf{11}$. Further, $p_{\max} = p_1 + p_2 + p_3$ and $P_{\max} = (p_1/p_{\max}, p_2/p_{\max}, p_3/p_{\max})$. Since $p_1 \geq p_2 \geq p_3$ it follows that

$$\frac{p_2 + p_3}{p_{\max}} \leq \frac{2}{3}.$$

This yields

$$\mathcal{L}_{\mathcal{C}_{\max}}^{1,2}(P_{\max}) = 1 + \frac{p_2 + p_3}{p_{\max}} \leq \frac{5}{3}.$$

It now follows from Equations (2.6) and (2.7) that

$$\mathcal{L}_{\mathcal{C}}^{1,2}(P) \leq 1 + \frac{5}{3} p_{\max} < 1 + \frac{5}{3} \cdot \frac{5}{8} = \frac{49}{24} < \frac{5}{2}.$$

For $t \geq 4$ the induction hypothesis and Equation (2.7) yield

$$\mathcal{L}_{\mathcal{C}}^{1,2}(P) < 1 + \frac{1 + 2(t-1)}{4(t-1)} \cdot \frac{5}{2} \leq 1 + \frac{7}{12} \cdot \frac{5}{2} = \frac{59}{24} < \frac{5}{2}.$$

Case 2.2.2: $\mathbf{v}_{\max} \in \mathcal{U}_1$

We get that $p_{\max} = \sum_{u=t+1}^{N} p_u$. If $\sum_{u=t+1}^{N} p_u \leq 1/2$, we get like before

$$\mathcal{L}_{\mathcal{C}}^{1,2}(P) < 1 + \frac{1}{2} \cdot \frac{5}{2} = \frac{9}{4} < \frac{5}{2}.$$

If $\sum_{u=t+1}^{N} p_u > 1/2$, it follows that

$$\sum_{u=1}^{t} p_u \geq \frac{1}{2} - \frac{1}{2} p_{t+1}.$$

Since $p_{t+1} \leq \left(\sum_{u=1}^{t} p_u\right)/t$, we further obtain

$$\sum_{u=1}^{t} p_u \geq \frac{t}{2t+1} \geq \frac{2}{5}. \tag{2.8}$$

Since $p_{\max} = 1 - \sum_{u=1}^{t} p_u$, we finally get by induction and Equation (2.8) that

$$\mathcal{L}_{\mathcal{C}}^{1,2}(P) < 1 + \frac{3}{5} \cdot \frac{5}{2} = \frac{5}{2}.$$

□

From Theorem 2 in [2] and Theorem 2.3 follows

Corollary 2.4 *It holds for all probability distributions P on $\mathcal{U}$ that*

$$2\left(1 - \sum_{u \in \mathcal{U}} p_u^2\right) \leq \mathcal{L}^{1,2}(P,P) \leq \mathcal{L}^{1,2}(P) < \frac{5}{2}.$$

3 L-Identification for the Uniform Distribution

In the first subsection we point out an interesting connection between the so-called *balanced Huffman codes for the uniform distribution* and the colexicographic order (see e.g. [14]). This order can be used to construct such codes. In the remaining we refer only to balanced Huffman codes and skip the add on "for the uniform distribution". This is somewhat detached from L-identification but since balanced Huffman codes are crucial for the analysis in the second subsection, we feel that this section is the right place to state this result.

We assume familiarity with the concept of Huffman coding (see [12]) and shall start by recalling the concept of balanced Huffman codes, which was introduced in [3]. Let $N = q^{n-1}+d$, where $0 \leq d \leq (q-1)q^{n-1}-1$. The q-ary Huffman coding for the uniform distribution of size N yields a code where some codewords have length n and the other codewords have length $n-1$. More precisely, if $0 \leq d < q^{n-1}$, then $q^{n-1}-d$ codewords have length $n-1$ and $2d$ codewords have length n, while in the case $q^{n-1} \leq d \leq (q-1)q^{n-1}-1$ all codewords have length n. It is well-known that for data compression all Huffman codes are optimal. This is not the case for identification.

In [3] it is shown (for $q = 2$) that for identification it is crucial which codewords have length n or, in terms of codetrees, where in the codetree these longer codewords lie. Moreover, those Huffman codes have a shorter expected and worst-case running time for which the longer codewords are distributed along the code tree in such a way that for every inner node the difference between the number of leaves of its left side and the number of leaves of its right side is at most one. In [3] Huffman trees satisfying this property were called *balanced*. By analogy, we shall also say that a q-ary Huffman code is balanced if its corresponding q-ary codetree $\mathcal{H}$ obeys the property that for every inner node $x \in \mathring{\mathcal{N}}(\mathcal{H})$ the difference between the number of leaves of $\mathcal{H}_{xi}$ and $\mathcal{H}_{xj}$ is at most one for all $i, j \in \mathcal{Q}$. We further denote by $\mathcal{H}_{q,N}$ the set of all q-ary balanced Huffman trees with N leaves and the corresponding set of q-ary balanced Huffman codes of size N is denoted by $\mathcal{C}_{q,N}$. If $N = q^n$, there exists only a single balanced Huffman code, namely $\mathcal{C}_{q^n}$. We denote the balanced Huffman tree which corresponds to $\mathcal{C}_{q^n}$ by $\mathcal{H}_{q^n}$.

In identification what is relevant is not the length of a codeword but the length of the maximal common prefix of two or more different codewords. This is why a balanced Huffman code is better for identification than an unbalanced one. It is easy to see by the pigeonhole principle that if we consider Huffman codes with codewords of lengths $n-1$ and n, a balanced Huffman code is optimal for the worst-case running time and we will see in the proof of Theorem 3.4 that the balancing property is also crucial for the symmetric running time of L-identification.

The q-ary Shannon-Fano coding procedure [10] constructs codes where for every inner node the difference between the sum of the normalized probabilities within its individual branches is as close as possible to $1/q$. It is an easy observation that if we are dealing with uniform distributions, a code is a Shannon-Fano code if and only if it is a balanced Huffman code.

The main result of this section is the examination of the asymptotic behavior of $\mathcal{L}_{\mathcal{C}}^{L,q}(P,P)$ for the case when P is the uniform distribution. We shall prove that this is equal to a rational number $K_{L,q}$ (Theorem 3.4), which grows logarithmically in L. In fact, we show that $K_{L,2}$ approximates the L-th harmonic number. We note that Theorem 3.4 also plays a major role in the discovery of the identification entropies, which are discussed in Sections 4 and 5.

3.1 Colexicographic Balanced Huffman Trees

In this subsection we will show how one can construct a balanced Huffman tree for given q, n and $N = q^{n-1} + d$ for some d by applying the colexicographic order. Therefore, let $k = \lfloor d/q^{n-1} \rfloor \leq q-2$ and $m = d \mod q^{n-1}$. Since a Huffman code contains only codewords of lengths $n-1$ and n, we begin our construction of a balanced Huffman tree with $\mathcal{H}_{q^{n-1}}$ and extend it into the next level by replacing all its leaves with copies of $\mathcal{H}_k$, which we call extension trees. We call this constructed tree the *base tree* $\mathcal{B}$. Obviously, $\mathcal{B}$ is a balanced Huffman tree. We still have m elements left which have to be inserted into the base tree. It remains to determine which ones of the extension trees will be used for this. Of course, every extension tree can only be used once, because otherwise the balancing property would be violated. Before we explain the construction which provides this, we formalize matters.

Let $A \subseteq \bar{\mathcal{N}}(\mathcal{H}_{q^{n-1}})$ be a set of leaves of $\mathcal{H}_{q^{n-1}}$. Then, we define $\mathcal{B}(A)$ to be the tree which we obtain by replacing all the extension trees of the base tree $\mathcal{B}$

with roots in $\mathcal{A}$ by $\mathcal{H}_{k+1}$. Such a set is called a *valid extension set*, if

$$\left| \; |\mathcal{B}(A)_{x_1 \ldots x_{\|x\|} i}| - |\mathcal{B}(A)_{x_1 \ldots x_{\|x\|} j}| \; \right| \leq 1 \tag{3.1}$$

for all $i, j \in \mathcal{Q}$ and all inner nodes $x \in \mathring{\mathcal{N}}(\mathcal{B})$. See Figure 3.1 for examples of a valid and an invalid extension. Equivalently we could have defined that A is a valid extension set if

$$| \; |A_{x,i}| - |A_{x,j}| \; | \leq 1 \tag{3.2}$$

for all $x \in \mathring{\mathcal{N}}(\mathcal{B})$ and all $i, j \in \mathcal{Q}$ and where $A_{x,i} = \{a \in A_x \; : \; a_{\|x\|+1} = i\}$ and $A_x = \{a \in A \; : \; a_1 ... a_{\|x\|} = x\}$. An immediate conclusion is that if A is a valid extension set, then $\mathcal{B}(A)$ is a balanced Huffman tree.

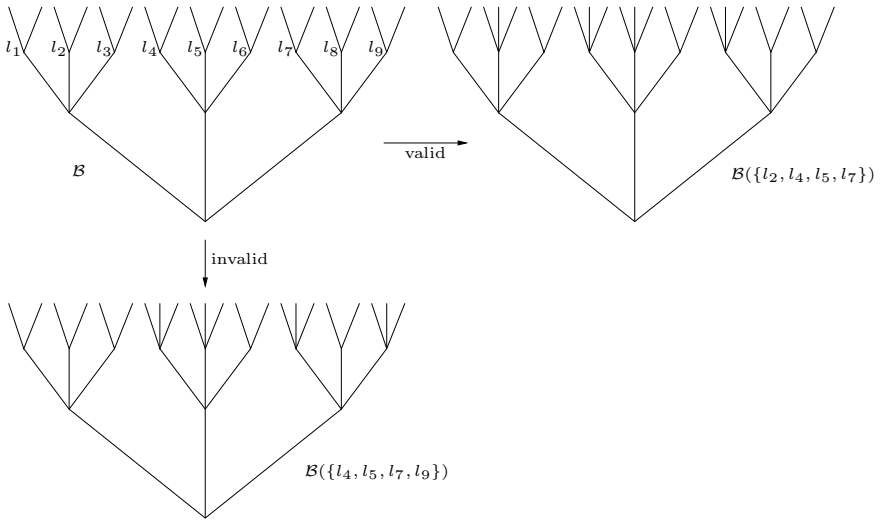

Figure 3.1: Examples for a valid and an invalid extension of the ternary base tree $\mathcal{B}$ for $N = 22$.

An easy consequence of the balancing property is the following

Lemma 3.1 *Let $q^{n-1} < N \leq q^n$, $\mathcal{H} \in \mathcal{H}_{q,N}$ and x be a node of $\mathcal{H}$, then it follows*

$$\left\lfloor \frac{N}{q^{\|x\|}} \right\rfloor \leq |\mathcal{H}_x| \leq \left\lceil \frac{N}{q^{\|x\|}} \right\rceil . \tag{3.3}$$

The inequality holds with equality for all x if and only if $N = q^n$. Moreover, it simplifies to

$$|\mathcal{H}_x| = q^{n-\|x\|}. \tag{3.4}$$

For given q and N there may exist many different balanced Huffman trees. We want to point out an interesting case the so-called *colexicographic* balanced Huffman tree. This tree is obtained by taking as the extension set A^{col} the first m codewords of length $n-1$ in colexicographic order.

Let $x, y \in \mathcal{Q}^{n-1}$ and $i_{\max} = \max\{i \in \{1,...,n-1\} \ : \ x_i \neq y_i\}$. Then x is said to be less or equal than y in the colexicographic order, denoted by $x \preceq y$, if $x_{i_{\max}} \leq y_{i_{\max}}$. One can easily verify that $(\mathcal{Q}^{n-1}, \preceq)$ is a linearly ordered set since $\mathcal{Q}^{n-1}$ is a product space and the colexicographic order is induced by the trivial linear $\leq$ order on $\mathcal{Q}$. If we denote by c_i the i-th codeword in this order and focus on the k-th q-bits, we observe the following structure.

$$c_{1,k}...c_{q^{n-1},k} = \underbrace{Q_k...Q_k}_{q^{n-k-1}},$$

where

$$Q_k = \underbrace{0...0}_{q^{k-1}}\underbrace{1...1}_{q^{k-1}}.....\underbrace{(q-1)...(q-1)}_{q^{k-1}}.$$

Moreover, the prefixes of length $k-1$ of the codewords within a block Q_k which coincide in the k-th q-bit form the complete $\mathcal{Q}^{k-1}$. And all the codewords in such a block have identical suffixes of length $n-k-1$.

We further define s_k and r_k by $m = s_k q^k + r_k$, where $r_k < q^k$ and $k \in [n-1]$. Finally, r'_k and r''_k are given by $r_k = r'_k q^{k-1} + r''_k$, where $0 \leq r''_k < q^{k-1}$. With this notation we obtain that the k-th q-bits of the first m codewords look like

$$c_{1,k}...c_{m,k} = \underbrace{Q_k...Q_k}_{s_k}\underbrace{0...0}_{q^{k-1}}...\underbrace{(r'_k-1)...(r'_k-1)}_{q^{k-1}}\underbrace{r'_k...r'_k}_{r''_k}.$$

Let $x \in \mathcal{B}$. With the notation of Equation (3.2) we get that A^{col}_x contains exactly q codewords from each of the s_k blocks Q_k each with a different k-th q-bit. In addition, it contains exactly one codeword from each of the small blocks $0...0$ to $(r'_k-1)...(r'_k-1)$ and at most one codeword from the partial small block $r'_k...r'_k$. This yields

$$|A^{\text{col}}_{x,i}| = \begin{cases} s_k + 1 & \text{if } i = 1,...,r'_k \\ s_k \text{ or } s_k + 1 & \text{if } i = r'_k + 1 \\ s_k & \text{if } i = r'_k + 2,...,q. \end{cases}$$

This together with Equation (3.2) shows that A_{col} is a valid extension set. For further information about linear orders see [14].

3.2 An Asymptotic Theorem

The goal of this subsection is to analyze the asymptotic behavior of

$$\mathcal{L}_{\mathcal{C}}^{L,q}\left((\frac{1}{N},...,\frac{1}{N}),(\frac{1}{N},...,\frac{1}{N})\right)=\frac{1}{N^{L+1}}\sum_{u_1,...,u_L,v=1}^{N}\mathcal{L}_{\mathcal{C}}^{L,q}(u^L,v), \tag{3.5}$$

with $\mathcal{C}\in\mathcal{C}_{q,N}$. This will be done by applying a different counting method. The above equation suggests to calculate $\mathcal{L}_{\mathcal{C}}^{L,q}(u^L,v)$ for all pairs (u^L,v) individually. Instead we merge all u^L having the same running time regarding some v into sets

$$\mathcal{R}_{\mathcal{C}}^{L,q}(k,v)=\left\{u^L\in\mathcal{U}^L\ :\ \mathcal{L}_{\mathcal{C}}^{L,q}(u^L,v)=k\right\} \tag{3.6}$$

for $k\in[\|c_v\|]$. The above defined sets also depend on N. As well as the L-identification functions in Subsection 1.2 they contain this dependency implicitly via $\mathcal{C}$. Equation (3.5) now becomes

$$\mathcal{L}_{\mathcal{C}}^{L,q}\left((\frac{1}{N},...,\frac{1}{N}),(\frac{1}{N},...,\frac{1}{N})\right)=\frac{1}{N^{L+1}}\sum_{v=1}^{N}\sum_{k=1}^{\|c_v\|}k|\mathcal{R}_{\mathcal{C}}^{L,q}(k,v)|. \tag{3.7}$$

In order to apply this equation we need to know upper and lower bounds on the cardinalities of these sets. Corollary 3.3 below provides such bounds and exact values for the case when N is a q-power. The base for this corollary is the following

Lemma 3.2 *Let $q^{n-1}<N\leq q^n$, $\mathcal{C}\in\mathcal{C}_{q,N}$, $\mathcal{H}=T_{\mathcal{C}}$ and $v\in\mathcal{U}$.*
Then, for $k\in[\|c_v\|-1]$ it holds that

$$|\mathcal{R}_{\mathcal{C}}^{L,q}(k,v)|=\sum_{m=1}^{L}\binom{L}{m}|\bar{\mathcal{N}}(\mathcal{H}_{c_v^{k-1}})\backslash\bar{\mathcal{N}}(\mathcal{H}_{c_v^{k}})|^m\left(N-|\bar{\mathcal{N}}(\mathcal{H}_{c_v^{k-1}})|\right)^{L-m}$$

and

$$|\mathcal{R}_{\mathcal{C}}^{L,q}(\|c_v\|,v)|=\sum_{m=1}^{L}\binom{L}{m}|\bar{\mathcal{N}}(\mathcal{H}_{c_v^{\|c_v\|-1}})|^m\left(N-|\bar{\mathcal{N}}(\mathcal{H}_{c_v^{\|c_v\|-1}})|\right)^{L-m}.$$

Proof:

In order to simplify notation we shall write $\mathcal{R}(k,v)$ for $\mathcal{R}_{\mathcal{C}}^{L,q}(k,v)$.

Case 1: $\mathbf{k = 1}$

The L-identification algorithm terminates after the first step if and only if the codewords of all components of u^L differ already in the first q-bit from $c_{v,1}$. This gives us

$$\mathcal{R}(1, v) = \left\{u^L \in [q^n]^L \ : \ c_{u_i} \in \bar{\mathcal{N}}(\mathcal{H})\backslash\bar{\mathcal{N}}(\mathcal{H}_{c_{v,1}}) \ \forall \ i \in [L]\right\}$$

and therewith

$$|\mathcal{R}(1, v)| = |\bar{\mathcal{N}}(\mathcal{H})\backslash\bar{\mathcal{N}}(\mathcal{H}_{c_{v,1}})|^L = (N - |\bar{\mathcal{N}}(\mathcal{H}_{c_{v,1}})|)^L.$$

This coincides with the first equation of Lemma 3.2.

Case 2: $\mathbf{k = 2, ..., \|c_v\| - 1}$

The identification time of u^L and v equals k if and only if it holds for all $i \in [L]$ that $c_{u_i}^k \neq c_v^k$ and that there exists at least one $i \in [L]$ such that $c_{u_i}^{k-1} = c_v^{k-1}$. This consideration yields

$$\begin{aligned}\mathcal{R}(k, v) = \Big\{u^L \in [q^n]^L \ : \quad & \exists \ i \in [L] \text{ with } c_{u_i} \in \bar{\mathcal{N}}(\mathcal{H}_{c_v^{k-1}})\backslash\bar{\mathcal{N}}(\mathcal{H}_{c_v^k}) \\ & \text{and } \ c_{u_i} \notin \bar{\mathcal{N}}(\mathcal{H}_{c_v^k}) \ \forall \ i \in [L]\Big\}.\end{aligned}$$

In order to count the elements we partition $\mathcal{R}(k, v)$ into L subsets $S_{k,1}, ..., S_{k,L}$, where

$$\begin{aligned}S_{k,m} = \Big\{u^L \in [q^n]^L \ : \quad & \exists \ i_1, ..., i_m \in [L] \text{ with } c_{u_{i_1}}, ..., c_{u_{i_m}} \in \bar{\mathcal{N}}(\mathcal{H}_{c_v^{k-1}})\backslash\bar{\mathcal{N}}(\mathcal{H}_{c_v^k}) \\ & \text{and } \ c_{u_i} \in \bar{\mathcal{N}}(\mathcal{H})\backslash\bar{\mathcal{N}}(\mathcal{H}_{c_v^{k-1}}) \ \forall \ i \in [L]\backslash\{i_1, ..., i_m\}\Big\}.\end{aligned}$$

If we fix the positions $i_1, ..., i_m$, we see that the number of possible vectors is

$$|\bar{\mathcal{N}}(\mathcal{H}_{c_v^{k-1}})\backslash\bar{\mathcal{N}}(\mathcal{H}_{c_v^k})|^m (N - |\bar{\mathcal{N}}(\mathcal{H}_{c_v^{k-1}})|)^{L-m}.$$

Since we have no restrictions for these positions, it follows that

$$|S_{k,m}| = \binom{L}{m} |\bar{\mathcal{N}}(\mathcal{H}_{c_v^{k-1}})\backslash\bar{\mathcal{N}}(\mathcal{H}_{c_v^k})|^m \left(N - |\bar{\mathcal{N}}(\mathcal{H}_{c_v^{k-1}})|\right)^{L-m}.$$

Altogether we obtain

$$\begin{aligned}|\mathcal{R}(k, v)| \quad &= |\bigcup_{m=1}^{L} S_{k,m}| = \sum_{m=1}^{L} |S_{k,m}| \\ &= \sum_{m=1}^{L} \tbinom{L}{m} |\bar{\mathcal{N}}(\mathcal{H}_{c_v^{k-1}})\backslash\bar{\mathcal{N}}(\mathcal{H}_{c_v^k})|^m \left(N - |\bar{\mathcal{N}}(\mathcal{H}_{c_v^{k-1}})|\right)^{L-m}.\end{aligned}$$

Case 3: $\mathbf{k = \|c_v\|}$

In this case also c_v itself may be one of the components of u^L. This yields

$$\mathcal{R}(n,v) = \{u^L \in [q^n]^L \ : \ \exists\, i \in [L] \text{ with } c_{u_i} \in \bar{\mathcal{N}}(\mathcal{H}_{c_v^{\|c_v\|-1}})\}.$$

According to this we adjust the subsets $S_{n,1}, ..., S_{n,L}$, such that

$$S_{n,m} = \{u^L \in [q^n]^L \ : \ \ \exists\, i_1, ..., i_m \in [L] \text{ with } c_{u_{i_1}}, ..., c_{u_{i_m}} \in \bar{\mathcal{N}}(\mathcal{H}_{c_v^{\|c_v\|-1}})$$
$$\text{and } c_{u_i} \in \bar{\mathcal{N}}(\mathcal{H}) \backslash \bar{\mathcal{N}}(\mathcal{H}_{c_v^{\|c_v\|-1}}) \ \forall\ i \in [L] \backslash \{i_1, ..., i_m\}\}.$$

Of course, these sets partition $\mathcal{R}(n,1)$ and since

$$|S_{n,m}| = \binom{L}{m} |\bar{\mathcal{N}}(\mathcal{H}_{c_v^{\|c_v\|-1}})|^m (N - |\bar{\mathcal{N}}(\mathcal{H}_{c_v^{\|c_v\|-1}})|)^{L-m},$$

for all $m \in [L]$, we obtain the desired result for $|\mathcal{R}(n,v)|$.

□

If we combine Lemma 3.1 and Lemma 3.2, we obtain

Corollary 3.3 *With the same definitions as in Lemma 3.2 we have the following upper bounds for $k \in [\|c_v\| - 1]$*

$$|\mathcal{R}_{\mathcal{C}}^{L,q}(k,v)| \leq \sum_{m=1}^{L} \binom{L}{m} \left(\left\lceil \frac{N}{q^{k-1}} \right\rceil - \left\lfloor \frac{N}{q^k} \right\rfloor \right)^m \left(N - \left\lfloor \frac{N}{q^{k-1}} \right\rfloor \right)^{L-m}$$

and

$$|\mathcal{R}_{\mathcal{C}}^{L,q}(\|c_v\|,v)| \leq \sum_{m=1}^{L} \binom{L}{m} \left\lceil \frac{N}{q^{\|c_v\|-1}} \right\rceil^m \left(N - \left\lfloor \frac{N}{q^{\|c_v\|-1}} \right\rfloor \right)^{L-m}.$$

Additionally, we get lower bounds for $k \in [\|c_v\| - 1]$

$$|\mathcal{R}_{\mathcal{C}}^{L,q}(k,v)| \geq \sum_{m=1}^{L} \binom{L}{m} \left(\left\lfloor \frac{N}{q^{k-1}} \right\rfloor - \left\lceil \frac{N}{q^k} \right\rceil \right)^m \left(N - \left\lceil \frac{N}{q^{k-1}} \right\rceil \right)^{L-m}$$

and

$$|\mathcal{R}_{\mathcal{C}}^{L,q}(\|c_v\|,v)| \geq \sum_{m=1}^{L} \binom{L}{m} \left\lfloor \frac{N}{q^{\|c_v\|-1}} \right\rfloor^m \left(N - \left\lceil \frac{N}{q^{\|c_v\|-1}} \right\rceil \right)^{L-m}.$$

The above inequalities hold with equality for all $v \in \mathcal{U}$ if and only if $N = q^n$. Moreover, they simplify for all $k \in [n-1]$ to

$$|\mathcal{R}_{\mathcal{C}}^{L,q}(k,v)| = q^{nL} \sum_{m=1}^{L} \binom{L}{m} q^{-km}(q-1)^m(1-q^{-k+1})^{L-m}$$

and

$$|\mathcal{R}_{\mathcal{C}}^{L,q}(\|c_v\|,v)| = \sum_{m=1}^{L} \binom{L}{m} q^m(q^n-q)^{L-m}.$$

With the above estimates we are now ready to prove the asymptotic theorem for uniform distributions. If we consider the uniform distribution and use a balanced Huffman code for the encoding, the symmetric L-identification running time asymptotically equals a rational number $K_{L,q}$.

Theorem 3.4 *Let $L, n \in \mathbb{N}$, $q \in \mathbb{N}_{\geq 2}$, $q^{n-1} < N \leq q^n$, $\mathcal{C} \in \mathcal{C}_{q,N}$ and P be the uniform distribution on $[N]$. Then it holds that*

$$\lim_{N\to\infty} \mathcal{L}_{\mathcal{C}}^{L,q}(P,P) = K_{L,q} = -\sum_{l=1}^{L} (-1)^l \tbinom{L}{l} \tfrac{q^l}{q^l-1}.$$

Proof:

Case 1: $\mathbf{N = q^n}$

It follows from Corollary 3.3 and Equation (3.7) that

$$\begin{aligned} \mathcal{L}_{\mathcal{C}}^{L,q}(P,P) = \tfrac{1}{q^{nL}} \quad \Bigg[\sum_{k=1}^{n-1} \quad & kq^{nL} \sum_{m=1}^{L} \tbinom{L}{m} q^{-km}(q-1)^m(1-q^{-k+1})^{L-m} \\ + \quad & n \sum_{m=1}^{L} \tbinom{L}{m} q^m(q^n-q)^{L-m} \Bigg]. \end{aligned} \tag{3.8}$$

It is easy to check that the second summand together with the leading factor q^{-nL} converges to 0 if n goes to infinity. In fact,

$$\sum_{m=1}^{L} \binom{L}{m} nq^{-m(n-1)}(1-q^{-n+1})^{L-m} \to 0.$$

This is because $nq^{-m(n-1)} \to 0$ and $(1-q^{-n+1})^{L-m} \to 1$. Thus, we get

$$\begin{aligned}\lim_{n\to\infty} \mathcal{L}_{\mathcal{C}}^{L,q}(P,P) &= \sum_{k=1}^{\infty} k \sum_{m=1}^{L} \tbinom{L}{m} q^{-km}(q-1)^m(1-q^{-k+1})^{L-m} \\ &= \sum_{m=1}^{L} \sum_{t=0}^{L-m} (-q)^t \tbinom{L}{m} \tbinom{L-m}{t} (q-1)^m \sum_{k=1}^{\infty} k q^{-k(m+t)} \qquad (3.9) \\ &= \sum_{m=1}^{L} \sum_{t=0}^{L-m} (-q)^t \tbinom{L}{m} \tbinom{L-m}{t} (q-1)^m \tfrac{q^{m+t}}{(q^{m+t}-1)^2}.\end{aligned}$$

The second equality follows from $(1-q^{-k+1})^{L-m} = \sum_{t=0}^{L-m} \binom{L-m}{t}(-q)^t q^{-tk}$, while the last equality is a consequence of the geometric series.

In the following we set $x_{m,t} = (-q)^t \binom{L}{m}\binom{L-m}{t}(q-1)^m$ as well as $z_l = q^l/(q^l-1)^2$ and change the order of summation. This yields

$$\begin{aligned}\lim_{n\to\infty} \mathcal{L}_{\mathcal{C}}^{L,q}(P,P) &= \sum_{m=1}^{L} \sum_{t=0}^{L-m} x_{m,t} z_{m+t} = \sum_{l=1}^{L} z_l \sum_{t=0}^{l-1} x_{l-t,t} \\ &= \sum_{l=1}^{L} \tfrac{q^l}{(q^l-1)^2} \sum_{t=0}^{l-1} (-q)^t \tbinom{L}{l-t} \tbinom{L-l+t}{t} (q-1)^{l-t} \\ &= \sum_{l=1}^{L} \tbinom{L}{l} \tfrac{q^l}{(q^l-1)^2} \sum_{t=0}^{l-1} \tbinom{l}{t} (-q)^t (q-1)^{l-t} \\ &= \sum_{l=1}^{L} \tbinom{L}{l} \tfrac{q^l}{(q^l-1)^2} \left((-1)^l - (-q)^l\right) \\ &= -\sum_{l=1}^{L} (-1)^l \tbinom{L}{l} \tfrac{q^l}{q^l-1}.\end{aligned}$$

Case 2: $\mathbf{q^{n-1} < N < q^n}$

For this case we obtain

$$\begin{aligned}
&\mathcal{L}_{\mathcal{C}}^{L,q}(P,P)\\
\leq\ & \tfrac{1}{N^{L+1}} \sum_{v=1}^{N} \left[\sum_{k=1}^{\|c_v\|-1} k \sum_{m=1}^{L} \binom{L}{m} \left(\lceil \tfrac{N}{q^{k-1}} \rceil - \lfloor \tfrac{N}{q^k} \rfloor \right)^m \left(N - \lfloor \tfrac{N}{q^{k-1}} \rfloor \right)^{L-m} \right]\\
&+ \tfrac{1}{N^{L+1}} \sum_{v=1}^{N} \left[\|c_v\| \sum_{m=1}^{L} \binom{L}{m} \lceil \tfrac{N}{q^{\|c_v\|-1}} \rceil^m (N - \lfloor \tfrac{N}{q^{\|c_v\|-1}} \rfloor)^{L-m} \right] \qquad (3.10)\\
\leq\ & \tfrac{1}{N} \sum_{v=1}^{N} \left[\sum_{k=1}^{\|c_v\|-1} k \sum_{m=1}^{L} \binom{L}{m} (q^{-k+1} - q^{-k} + \tfrac{2}{N})^m (1 - q^{-k+1} + \tfrac{1}{N})^{L-m} \right]\\
&+ \tfrac{1}{N} \sum_{v=1}^{N} \left[\|c_v\| \sum_{m=1}^{L} \binom{L}{m} (q^{-\|c_v\|+1} + \tfrac{1}{N})^m (1 - q^{-\|c_v\|+1} + \tfrac{1}{N})^{L-m} \right].
\end{aligned}$$

The first inequality is obtained by the insertion of the upper bound in Corollary 3.3 into Equation (3.7). $\lceil N/q^k \rceil \leq N/q^k + 1$ and $\lfloor N/q^k \rfloor \geq N/q^k - 1$ yield the second inequality. We now divide this case into two subcases.

Case 2.1: $\mathbf{2q^{n-1} \leq N < q^n}$

In this case all codewords have length n. Hence Equation 3.10 reduces to

$$\begin{aligned}
\mathcal{L}_{\mathcal{C}}^{L,q}(P,P) \leq\ & \sum_{k=1}^{n-1} k \sum_{m=1}^{L} \binom{L}{m} (q^{-k+1} - q^{-k} + \tfrac{2}{N})^m (1 - q^{-k+1} + \tfrac{1}{N})^{L-m}\\
&+ n \sum_{m=1}^{L} \binom{L}{m} (q^{-n+1} + \tfrac{1}{N})^m (1 - q^{-n+1} + \tfrac{1}{N})^{L-m}.
\end{aligned} \qquad (3.11)$$

As in the case $N = q^n$ the second summand goes to zero as N goes to infinity. Thus, we only have to consider the first summand. In fact, we can reduce this case to the previous one by applying the binomial theorem. We obtain

$$\left(q^{-k}(q-1) + \frac{2}{N} \right)^m = \left(q^{-k}(q-1) \right)^m + \sum_{t=0}^{m-1} \binom{m}{t} (q^{-k}(q-1))^t \left(\frac{2}{N} \right)^{m-t}$$

and

$$\left(1 - q^{-k+1} + \frac{1}{N} \right)^{L-m} = \left(1 - q^{-k+1} \right)^{L-m} + \sum_{s=0}^{L-m-1} \binom{L-m}{s} (1-q^{-k+1})^s \frac{1}{N^{L-m-s}}.$$

In the following we use

$$A = \sum_{t=0}^{m-1} \binom{m}{t} (q^{-k}(q-1))^t \left(\frac{2}{N}\right)^{m-t}$$

and

$$B = \sum_{s=0}^{L-m-1} \binom{L-m}{s} (1-q^{-k+1})^s \frac{1}{N^{L-m-s}}.$$

With this notation the right hand side of Equation (3.11) asymptotically equals

$$\begin{aligned} \sum_{k=1}^{n-1} k \sum_{m=1}^{L} \tbinom{L}{m} \Big[\left(q^{-k}(q-1)\right)^m \left(1-q^{-k+1}\right)^{L-m} + \left(q^{-k}(q-1)\right)^m B \\ + \left(1-q^{-k+1}\right)^{L-m} A + AB \Big]. \end{aligned} \tag{3.12}$$

If we focus on the second summand in the square brackets, we see that

$$\begin{aligned} & \sum_{k=1}^{n-1} k \sum_{m=1}^{L} \tbinom{L}{m} \left(q^{-k}(q-1)\right)^m B \\ = & \sum_{m=1}^{L} \sum_{s=0}^{L-m-1} \tbinom{L-m}{s} \tbinom{L}{m} \tfrac{(q-1)^m}{N^{L-m-s}} \sum_{k=1}^{n-1} k q^{-km} \left(1-q^{-k+1}\right)^{L-m} \\ = & \sum_{m=1}^{L} \sum_{s=0}^{L-m-1} \sum_{r=0}^{L-m} (-1)^r \tbinom{L-m}{r} \tbinom{L-m}{s} \tbinom{L}{m} \tfrac{q^r(q-1)^m}{N^{L-m-s}} \sum_{k=1}^{n-1} k q^{-k(m+r)} \\ = & \sum_{m=1}^{L} \sum_{s=0}^{L-m-1} \sum_{r=0}^{L-m} \alpha(m,s,r) \tfrac{1}{N^{L-m-s}} \tfrac{1}{(q^{m+r}-1)^2} \left(q^{m+r} - \tfrac{(q^{m+r}-1)n+q^{m+r}}{q^{n(m+r)}}\right), \end{aligned}$$

where $\alpha(m,s,r) = (-1)^r \binom{L-m}{r}\binom{L-m}{s}\binom{L}{m} q^r (q-1)^m$. The last equality follows from the partial sum behavior of the geometric series. This expression tends to zero as N (resp. $n \approx \log_q N$) goes to infinity because $L-m-s \geq 1$.

In the same way it can be shown that the third and the fourth summand of Equation (3.12) also tend to zero. Thus, we end up with exactly the same expression like Equation (3.9). This proves the upper bound for this case. By using the same arguments and the lower estimates in Corollary 3.3 one can easily show the matching lower bound.

Case 2.2: $\mathbf{q^{n-1} < N < 2q^{n-1}}$

In this case $N = q^{n-1} + d$, with $0 < d < q^{n-1}$, and there exist exactly $q^{n-1} - d$ codewords of length $n-1$ and $2d$ codewords of length n. Then, Equation (3.10) becomes

$$
\begin{aligned}
& \mathcal{L}_{\mathcal{C}}^{L,q}(P,P) \\
\leq\ & \tfrac{q^{n-1}-d}{N} \Bigg[\sum_{k=1}^{n-2} k \sum_{m=1}^{L} \tbinom{L}{m} (q^{-k+1} - q^{-k} + \tfrac{2}{N})^m (1 - q^{-k+1} + \tfrac{1}{N})^{L-m} \\
& \qquad + (n-1) \sum_{m=1}^{L} \tbinom{L}{m} (q^{-n+2} + \tfrac{1}{N})^m (1 - q^{-n+2} + \tfrac{1}{N})^{L-m} \Bigg] \\
& + \tfrac{2d}{N} \Bigg[\sum_{k=1}^{n-1} k \sum_{m=1}^{L} \tbinom{L}{m} (q^{-k+1} - q^{-k} + \tfrac{2}{N})^m (1 - q^{-k+1} + \tfrac{1}{N})^{L-m} \\
& \qquad + n \sum_{m=1}^{L} \tbinom{L}{m} (q^{-n+1} + \tfrac{1}{N})^m (1 - q^{-n+1} + \tfrac{1}{N})^{L-m} \Bigg] \\
=\ & \sum_{k=1}^{n-2} k \sum_{m=1}^{L} \tbinom{L}{m} (q^{-k+1} - q^{-k} + \tfrac{2}{N})^m (1 - q^{-k+1} + \tfrac{1}{N})^{L-m} \\
& + \tfrac{(q^{n-1}-d)(n-1)}{N} \sum_{m=1}^{L} \tbinom{L}{m} (q^{-n+2} + \tfrac{1}{N})^m (1 - q^{-n+2} + \tfrac{1}{N})^{L-m} \\
& + \tfrac{2d(n-1)}{N} \sum_{m=1}^{L} \tbinom{L}{m} (q^{-n+2} - q^{-n+1} + \tfrac{2}{N})^m (1 - q^{-n+2} + \tfrac{1}{N})^{L-m} \\
& + \tfrac{2dn}{N} \sum_{m=1}^{L} \tbinom{L}{m} (q^{-n+1} + \tfrac{1}{N})^m (1 - q^{-n+1} + \tfrac{1}{N})^{L-m}.
\end{aligned}
$$

For the same reason as in the preceding cases the last three summands tend to zero as $N \to \infty$ and since the first summand asymptotically equals the first summand of Equation (3.11), the upper bound also in this last case is settled. Omitting the details we limit ourselves to remark that also in this case the matching lower bound can be easily obtained by a perfectly analogous argument. Thus, the proof of the theorem is complete.

□

A natural question regards the asymptotic growth of $K_{L,q}$ with respect to L. Table 3.1 shows some values of $K_{L,2}$. This motivates the assumption that $K_{L,2}$ grows logarithmically in L. In fact, this assumption proves true by the following considerations. First, we see that

$$K_{L,2} = -\sum_{l=1}^{L}(-1)^l \binom{L}{l} \frac{2^l}{2^l - 1} = 1 - \sum_{l=1}^{L} \frac{(-1)^l \binom{L}{l}}{2^l - 1}.$$

By using the geometric series we get

$$K_{L,2} - 1 = -\sum_{l=1}^{L} \frac{(-1)^l \binom{L}{l}}{2^l} \sum_{k=0}^{\infty} 2^{-kl} = -\sum_{k=0}^{\infty}\sum_{l=1}^{L} \binom{L}{l} (-1)^l 2^{-(k+1)l}.$$

The binomial theorem now yields

$$K_{L,2} - 1 = -\sum_{k=0}^{\infty}((1 - 2^{-(k+1)})^L - 1) = \sum_{k=1}^{\infty}(1 - (1 - 2^{-k})^L).$$

If we now set $\xi_k = (1 - 2^{-k})$, we obtain

$$\begin{aligned} K_{L,2} - 1 = \sum_{k=1}^{\infty}(1 - \xi_k^L) &= \sum_{k=1}^{\infty}(1 - \xi_k)(1 + \xi_k + \xi_k^2 + ... + \xi_k^{L-1}) \\ &= \sum_{k=1}^{\infty} \tfrac{1}{2^k}(1 + \xi_k + \xi_k^2 + ... + \xi_k^{L-1}). \end{aligned}$$

Figure 3.2 shows that this expression is an approximation by the upper sum of the integral

$$\int_0^1 (1 + x + x^2 + ... + x^{L-1})dx = 1 + \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{L} = H_L,$$

where H_L denotes the L-th harmonic number. Since H_L grows logarithmically with respect to L, so does $K_{L,2}$.

Volker Strehl ([16]) generalized this result for the case $q > 2$. His result is the content of the following

Proposition 3.5 (V. Strehl [16])
It holds that

$$\lim_{L\to\infty} \frac{H_L}{K_{L,q}} = \ln q,$$

where H_L denotes the L-th harmonic number and ln *is the natural logarithm.*

L	1	2	2^2	2^3	2^5	2^{10}	2^{13}
$K_{L,2} \approx$	2	$2,6667$	$3,5048$	$4,4211$	$6,3552$	$11,3335$	$14,3328$
$\frac{K_{L,2}-1}{\log L} \approx$	*	$1,6667$	$1,2524$	$1,1404$	$1,0710$	$1,0333$	$1,0256$

Table 3.1: The growth of $K_{L,2}$ in L.

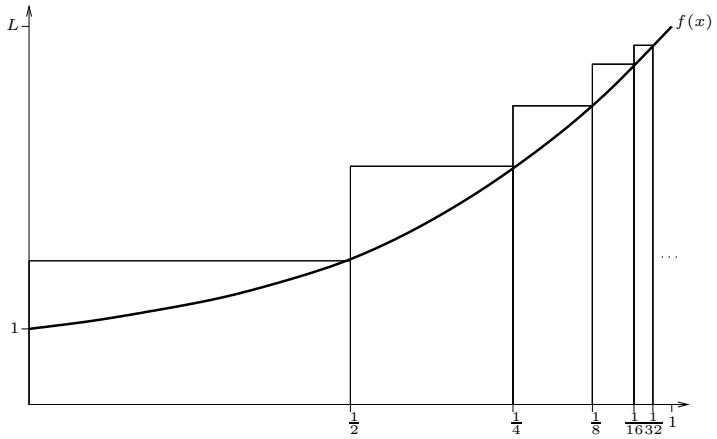

Figure 3.2: $K_{L,2} - 1$ approximates the integral of $f(x) = 1 + x + x^2 + ... + x^{L-1}$.

4 2-Identification for General Distributions

In the previous section we have seen how L-identification behaves for the uniform distribution. In this section we turn to general distributions and establish a lower bound for 2-identification.

Let us focus on the case $L = 2$, $N = q^n$, $P = (1/q^n, ..., 1/q^n)$ and $\mathcal{C} = \mathcal{C}_{q^n}$. Every q-ary comparison which is done during 2-identification for u^2 and v is itself an l-identification ($l \in [2]$) between the t-th q-bit of the codewords of the l still possible candidates and $c_{v,t}$. The running time of each of those "small" identifications is 1 no matter of the value of l. In fact, we have applied up to n "small" identifications within the code $\mathcal{C}_q$ in order to perform the original 2-identification within $\mathcal{C}_{q^n}$.

It is clear that $\mathcal{C}_{q^n} = \mathcal{C}_q^n$. Further, let $r_{t+1,l}$ be the probability that after the t-th comparison there are still l possible candidates left. We can now calculate 2-identification running time within $\mathcal{C}_q^n$ by

$$\begin{aligned} & \mathcal{L}^{2,q}_{\mathcal{C}_q^n}\left(\left(\tfrac{1}{q^n}, ..., \tfrac{1}{q^n}\right), \left(\tfrac{1}{q^n}, ..., \tfrac{1}{q^n}\right)\right) \\ = \;& 1 + \sum_{t=1}^{n-1} \sum_{l=1}^{2} \binom{2}{l} r_{t+1,l} \mathcal{L}^{l,q}_{\mathcal{C}_q}\left(\left(\tfrac{1}{q}, ..., \tfrac{1}{q}\right), \left(\tfrac{1}{q}, ..., \tfrac{1}{q}\right)\right) \\ = \;& 1 + 2\sum_{t=1}^{n-1} r_{t+1,1} + \sum_{t=1}^{n-1} r_{t+1,2}. \end{aligned}$$

Here, the binomial coefficient in the first equality occurs since in the case $l = 1$ either u_1 or u_2 is the leftover candidate. We have to take into account both possibilities. As stated before l-identification running time within $\mathcal{C}_q$ always equals 1. This proves the second equality. This approach yields an alternative proof of Theorem 3.4 for $L = 2$ and $|\mathcal{U}| = q^n$. However, we stop this analysis here and will come back to it later.

The above observations lead us to the attempt of doing the same for any given source code $\mathcal{C}$. Namely, to consider the discrete memoryless source $\left(\left(\mathcal{U}^n\right)^2, \left(P^n\right)^2\right)$

together with the concatenated code $\mathcal{C}^n$ and try to establish a connection between the 2-identification running time within $\mathcal{C}^n$ and the l-identification running times within $\mathcal{C}$. This relation is the content of Lemma 4.1. It turns out that we also have to consider (1-)identification within the basic code. This fact makes further analysis more sophisticated, especially for the general case of Section 5.

In order to apply Theorem 3.4 we firstly let n go to infinity. The result of this procedure is stated in Corollary 4.2. It is a consequence of Lemma 4.1. Furthermore, we show that from a particular concatenation step on we can lower bound all further concatenated codes to a saturated code $\mathcal{C}_{q^{K_n}}$ of some given length K_n. This is done in the proof of Lemma 4.5. Finally, Corollary 4.4 states that the uniform distribution is optimal for 2-identification within a block code. Altogether at the end of the first subsection we obtain

$$\mathcal{L}_{\mathcal{C}}^{2,q}(P,P) \geq (1-\sum_{u\in\mathcal{U}} p_u^3)\left(2\frac{q}{q-1}-\frac{q^2}{q^2-1}\right) - 2\left(\frac{1-\sum\limits_{u\in\mathcal{U}} p_u^3}{1-\sum\limits_{u\in\mathcal{U}} p_u^2}-1\right)\mathcal{L}_{\mathcal{C}}^{1,q}(P,P).$$

as a lower bound for 2-identification.

Unfortunately, (1-)identification appears negatively signed so that we cannot immediately apply the lower bound $\mathcal{L}_{\mathcal{C}}^{1,q}(P,P) \geq H_{\mathrm{ID}}^{1,q}(P)$, which has been proven in [5]. In the same work it has been shown that this lower bound is attainable if P consists only of q-powers. Proposition 4.7 at the beginning of the second subsection proves this equality also for 2-identification. This is the base for the definition of the *q-ary identification entropy of second degree*

$$H_{\mathrm{ID}}^{2,q}(P) = 2\frac{q}{q-1}\left(1-\sum_{u\in\mathcal{U}} p_u^2\right) - \frac{q^2}{q^2-1}\left(1-\sum_{u\in\mathcal{U}} p_u^3\right).$$

In the remaining part of the second subsection we prove some fundamental properties of this function. There are symmetry, expansibility, normalization, decisiveness, bounding between 0 and the uniform distribution and a special grouping behavior. Using these properties we prove Theorem 4.8 where we show that $H_{\mathrm{ID}}^{2,q}(P)$ is a lower bound for 2-identification. We end this part with a corollary which states that if we consider the uniform distribution on $\mathcal{U}$, balanced Huffman codes are asymptotically optimal for 2-identification.

Finally, we establish an upper bound for the binary case in the third Subsection. The code construction in the proof coincides with the one used for (1)-identification in Subsection 2.2.

4.1 An Asymptotic Approach

Lemma 4.1 *Let $\mathcal{U}$ be a finite set, $q \in \mathbb{N}_{\geq 2}$, P be a probability distribution on $\mathcal{U}$ and $\mathcal{C}$ be a prefix code. It holds that*

$$\begin{aligned}\mathcal{L}^{2,q}_{\mathcal{C}^n}(P^n, P^n) &= \mathcal{L}^{2,q}_{\mathcal{C}}(P,P)\left(1+\sum_{t=1}^{n-1}\left(\sum_{u\in\mathcal{U}} p_u^3\right)^t\right)\\ &\quad +2\mathcal{L}^{1,q}_{\mathcal{C}}(P,P)\left(\sum_{t=1}^{n-1}\left(\sum_{u\in\mathcal{U}} p_u^2\right)^t - \sum_{t=1}^{n-1}\left(\sum_{u\in\mathcal{U}} p_u^3\right)^t\right).\end{aligned}$$

Proof:

It is clear that while we are in the first basic tree we have to apply 2-identification and there are three possibilities of what might happen.

1. Both elements u_1^n and u_2^n do not coincide with v^n.
   The reason would be that their first components $u_{1,1}, u_{2,1}$ do not coincide with v_1. This stops the identification process.

2. Only one element, e.g. u_1^n, coincides with v^n.
   This would be because $u_{1,1} = v_1$ and $u_{2,1} \neq v_1$. Then, we continue with applying (1-)identification in the next tree (resp. code).

3. Both elements coincide with v^n.
   In this case also in the next tree 2-identification would have to be applied.

The main idea now is to exploit the fact that the symmetric 2-identification running time is an expectation. Therefore we introduce X_{t+1} as the random variable which indicates how many components of (U_1^n, U_2^n) are still candidates at step t. For all $t \in \{1, ..., n-1\}$ we define

$$X_{t+1} = \begin{cases} 0 & \text{if } U_1^t \neq V^t \text{ and } U_2^t \neq V^t \\ 1 & \text{if } (U_1^t = V^t \text{ and } U_2^t \neq V^t) \text{ or } (U_1^t \neq V^t \text{ and } U_2^t = V^t) \\ 2 & \text{if } U_1^t = U_2^t = V^t \end{cases}$$

and we set $X_1 = 2$. In order to calculate the corresponding probabilities we use the facts that U_1, U_2 and V are independent identically distributed. With this we get

$$\begin{aligned}Prob(X_{t+1} = 2) &= Prob(U_1^t = U_2^t = V^t)\\ &= \sum_{u^t\in\mathcal{U}^t} p_{u^t}^3 = \sum_{u_1,...,u_t\in\mathcal{U}} (p_{u_1}...p_{u_t})^3 = \left(\sum_{u\in\mathcal{U}} p_u^3\right)^t\end{aligned}$$

and

$$\begin{aligned} Prob(X_{t+1}=1) &= 2Prob(U_1^t = V^t \text{ and } U_2^t \neq V^t) = 2\sum_{u^t \in \mathcal{U}^t} p_{u^t}^2(1-p_{u^t}) \\ &= 2\left[\sum_{u_1,\ldots,u_t \in \mathcal{U}} (p_{u_1}\ldots p_{u_t})^2 - \sum_{u_1,\ldots,u_t \in \mathcal{U}} (p_{u_1}\ldots p_{u_t})^3\right] \\ &= 2\left[\left(\sum_{u\in\mathcal{U}} p_u^2\right)^t - \left(\sum_{u\in\mathcal{U}} p_u^3\right)^t\right]. \end{aligned}$$

As stated before the symmetric 2-identification running time is an expectation. Since for the first timestep $X_1 = 2$ and for all other timesteps the case $X_t = 0$ leads to the termination of the identification process before timestep t, we obtain

$$\begin{aligned} \mathcal{L}_{\mathcal{C}^n}^{2,q}(P^n, P^n) &= \sum_{t=1}^{n} \mathbb{E}(\mathcal{L}_{\mathcal{C}}^{X_t,q}(P,P)) = \sum_{t=0}^{n-1} \mathbb{E}(\mathcal{L}_{\mathcal{C}}^{X_{t+1},q}(P,P)) \\ &= \mathcal{L}_{\mathcal{C}}^{2,q}(P,P) + \sum_{t=1}^{n-1} Prob(X_{t+1}=1)\mathcal{L}_{\mathcal{C}}^{1,q}(P,P) \\ &\quad + \sum_{t=1}^{n-1} Prob(X_{t+1}=2)\mathcal{L}_{\mathcal{C}}^{2,q}(P,P) \\ &= \mathcal{L}_{\mathcal{C}}^{2,q}(P,P)\left(1 + \sum_{t=1}^{n-1}\left(\sum_{u\in\mathcal{U}} p_u^3\right)^t\right) \\ &\quad + 2\mathcal{L}_{\mathcal{C}}^{1,q}(P,P)\left(\sum_{t=1}^{n-1}\left(\sum_{u\in\mathcal{U}} p_u^2\right)^t - \sum_{t=1}^{n-1}\left(\sum_{u\in\mathcal{U}} p_u^3\right)^t\right). \end{aligned}$$

□

If we now establish the limit for n going to infinity and apply the geometric series for $k = 2, 3$ we obtain

$$\sum_{t=1}^{\infty}\left(\sum_{u\in\mathcal{U}} p_u^k\right)^t = \frac{1}{1-\sum_{u\in\mathcal{U}} p_u^k} - 1$$

and thus,

$$\lim_{n\to\infty} \mathcal{L}_{\mathcal{C}^n}^{2,q}(P^n,P^n) = \frac{\mathcal{L}_{\mathcal{C}}^{2,q}(P,P)}{1-\sum\limits_{u\in\mathcal{U}} p_u^3} + 2\left(\frac{1}{1-\sum\limits_{u\in\mathcal{U}} p_u^2} - \frac{1}{1-\sum\limits_{u\in\mathcal{U}} p_u^3}\right)\mathcal{L}_{\mathcal{C}}^{1,q}(P,P).$$

This proves

Corollary 4.2 *Let $\mathcal{U}$ be a finite set, $q \in \mathbb{N}_{\geq 2}$, P be a probability distribution on $\mathcal{U}$ and $\mathcal{C}$ be prefix code. It then holds that*

$$\mathcal{L}_{\mathcal{C}}^{2,q}(P,P) = (1-\sum_{u\in\mathcal{U}} p_u^3) \lim_{n\to\infty} \mathcal{L}_{\mathcal{C}^n}^{2,q}(P^n,P^n) - 2\left(\frac{1-\sum_{u\in\mathcal{U}} p_u^3}{1-\sum_{u\in\mathcal{U}} p_u^2} - 1\right) \mathcal{L}_{\mathcal{C}}^{1,q}(P,P).$$

Let us go back to the case where $\mathcal{U} = [q]$, $P = (1/q, ..., 1/q)$ and $\mathcal{C} = \mathcal{C}_q$. In this case $\mathcal{L}_{\mathcal{C}}^{l,q}(P,P) = 1$ for $l \in [2]$. It follows immediately from Corollary 4.2 that

$$\lim_{n\to\infty} \mathcal{L}_{\mathcal{C}^n}^{2,q}(P^n,P^n) = 2\frac{q}{q-1} - \frac{q^2}{q^2-1}. \tag{4.1}$$

This is the promised alternative proof of Theorem 3.4 for $L = 2$ and $|\mathcal{U}| = q^n$.

What we do now is to lower bound the expression $\lim_{n\to\infty} \mathcal{L}_{\mathcal{C}^n}^{2,q}(P^n,P^n)$. In Lemma 4.5 we show that we can limit ourselves to typical sequences (see [9]). Then we cut the codetree at some given depth and fill up the shorter branches to this depth with zero probability elements in order to obtain a saturated tree, resp. a block code. This does not increase the symmetric identification running time.

In Theorem 3.4 of Section 3, we have shown how L-identification and in particular 2-identification behaves asymptotically on block codes if we consider the uniform distribution. To use this result we have to show that for 2-identification uniform distribution is optimal for block codes. The following lemma provides a way for coming from any probability distribution to the uniform distribution without increasing the symmetric identification running time.

Lemma 4.3 *Let $n \in \mathbb{N}$, $q \in \mathbb{N}_{\geq 2}$, $k \in \{0, ..., n-1\}$ and $t \in \{0, ..., q^{n-k-1}-1\}$. Further, let $P = (p_1, ..., p_{q^n})$ and $\tilde{P} = (\tilde{p}_1, ..., \tilde{p}_{q^n})$ be probability distributions on $[q^n]$ with*

$$P = (p_1, ..., p_{tq^{k+1}}, \underbrace{r_1, ..., r_1}_{q^k}, \underbrace{r_2, ..., r_2}_{q^k}, ..., \underbrace{r_q, ..., r_q}_{q^k}, p_{(t+1)q^{k+1}+1}, ..., p_{q^n})$$

and

$$\tilde{P} = (p_1, ..., p_{tq^{k+1}}, \underbrace{\frac{1}{q}\sum_{i=1}^{q} r_i, ..., \frac{1}{q}\sum_{i=1}^{q} r_i}_{q^{k+1}}, p_{(t+1)q^{k+1}+1}, ..., p_{q^n}).$$

Then it holds that

$$\mathcal{L}^{2,q}_{\mathcal{C}_{q^n}}(P,P) - \mathcal{L}^{2,q}_{\mathcal{C}_{q^n}}(\tilde{P},\tilde{P}) \geq 0.$$

The inequality holds with equality if and only if either $k = 0$ or $r_i = r_j$ for all $i, j \in [q]$.

Proof:

W.l.o.g. we further assume that $t = 0$ such that for $i \in [q]$

$$p_{(i-1)q^k+1} = p_{(i-1)q^k+2} = ... = p_{iq^k} = r_i.$$

Also, we use for simplicity the abbreviations $L_{u_1u_2,v} = \mathcal{L}^{2,q}_{\mathcal{C}_{q^n}}((u_1,u_2),v)$ and $\alpha_{u_1u_2,v} = (p_{u_1}p_{u_2}p_v - \tilde{p}_{u_1}\tilde{p}_{u_2}\tilde{p}_v)L_{u_1u_2,v}$. With this notation we obtain

$$\begin{aligned}
& \mathcal{L}^{L,q}_{\mathcal{C}_{q^n}}(P,P) - \mathcal{L}^{L,q}_{\mathcal{C}_{q^n}}(\tilde{P},\tilde{P}) \\
= & \sum_{u_1,u_2,v=1}^{q^n} \alpha_{u_1u_2,v} \\
= & \sum_{v=1}^{q^n}\left[\sum_{u_1,u_2=1}^{q^{k+1}} \alpha_{u_1u_2,v} + 2\sum_{u_1=1}^{q^{k+1}}\sum_{u_2=q^{k+1}+1}^{q^n} \alpha_{u_1u_2,v} + \sum_{u_1,u_2=q^{k+1}+1}^{q^n} \alpha_{u_1u_2,v}\right] \\
= & \sum_{i=1}^{6} R_i,
\end{aligned} \tag{4.2}$$

where the second equality comes from the fact that $L_{u_1u_2,v} = L_{u_2u_1,v}$ and where

$$R_1 = \sum_{u_1,u_2,v=1}^{q^{k+1}} \alpha_{u_1u_2,v} \qquad R_2 = \sum_{u_1,u_2=1}^{q^{k+1}}\sum_{v=q^{k+1}+1}^{q^n} \alpha_{u_1u_2,v}$$

$$R_3 = 2\sum_{u_1,v=1}^{q^{k+1}}\sum_{u_2=q^{k+1}+1}^{q^n} \alpha_{u_1u_2,v} \qquad R_4 = 2\sum_{u_1=1}^{q^{k+1}}\sum_{u_2,v=q^{k+1}+1}^{q^n} \alpha_{u_1u_2,v}$$

$$R_5 = \sum_{u_1,u_2=q^{k+1}+1}^{q^n}\sum_{v=1}^{q^{k+1}} \alpha_{u_1u_2,v} \qquad R_6 = \sum_{u_1,u_2,v=q^{k+1}+1}^{q^n} \alpha_{u_1u_2,v}.$$

As one might expect the above summands disappear, except for R_1 and R_3. This is obvious for R_6 since $p_u = \tilde{p}_u$ for all $u \in [q^{k+1}+1, q^n]$.

If $u_1, u_2 \in [q^{k+1}+1, q^n]$, we have on the one hand that $L_{u_1u_2,v} = L_{u_1u_2,1}$ for all $v \in [q^{k+1}]$. We denote this by $L_{u_1u_2}$. On the other hand $p_{u_i} = \tilde{p}_{u_i}$ for $i = 1, 2$.

This yields

$$\begin{aligned} R_5 &= \sum_{u_1,u_2=q^{k+1}+1}^{q^n} \sum_{v=1}^{q^{k+1}} L_{u_1u_2} p_{u_1} p_{u_2} \left[p_v - \tfrac{1}{q} \sum_{i=1}^{q} r_i \right] \\ &= \sum_{u_1,u_2=q^{k+1}+1}^{q^n} L_{u_1u_2} p_{u_1} p_{u_2} \left[\sum_{v=1}^{q^{k+1}} p_v - q^k \sum_{i=1}^{q} r_i \right] = 0. \end{aligned}$$

Here, the final equality follows from $\sum_{v=1}^{q^{k+1}} p_v = \sum_{i=1}^{q} q^k r_i$.

If $u_2, v \in [q^{k+1}+1, q^n]$ and $u_1 \in [q^{k+1}]$, we see that $L_{u_1u_2,v} = L_{1u_2,v}$ and $p_{u_2} = \tilde{p}_{u_2}$ as well as $p_v = \tilde{p}_v$. Thus, proceeding as before we have that $R_4 = 0$.

If $u_1, u_2 \in [q^{k+1}]$ and $v \in [q^{k+1}+1, q^n]$, it follows that $L_{u_1u_2,v} = L_{11,v}$, which is denoted by L_v, and $p_v = \tilde{p}_v$. With this we get

$$\begin{aligned} R_2 &= \sum_{u_1,u_2=1}^{q^{k+1}} \sum_{v=q^{k+1}+1}^{q^n} L_v p_v \left[p_{u_1} p_{u_2} - \tfrac{1}{q^2} \left(\sum_{i=1}^{q} r_i \right)^2 \right] \\ &= \sum_{v=q^{k+1}+1}^{q^n} L_v p_v \left[\sum_{u_1,u_2=1}^{q^{k+1}} p_{u_1} p_{u_2} - q^{2k} \left(\sum_{i=1}^{q} r_i \right)^2 \right] = 0. \end{aligned}$$

Here, $\sum_{u_1,u_2=1}^{q^{k+1}} p_{u_1} p_{u_2} = \left(\sum_{i=1}^{q} q^k r_i \right)^2$ yields the final equality. Altogether we end up with

$$\mathcal{L}^{L,q}_{\mathcal{C}_{q^n}}(P,P) - \mathcal{L}^{L,q}_{\mathcal{C}_{q^n}}(\tilde{P},\tilde{P}) = R_1 + R_3.$$

We begin our remaining examinations with R_3. Similar as before we get $L_{u_1u_2,v} = L_{u_11,v}$, which we denote by $L_{u_1,v}$, and $p_{u_2} = \tilde{p}_{u_2}$ if $u_1, v \in [q^{k+1}]$ and $u_2 \in [q^{k+1}+1, q^n]$. We obtain

$$\begin{aligned} \tfrac{1}{2} R_3 &= \sum_{u_1,v=1}^{q^{k+1}} \sum_{u_2=q^{k+1}+1}^{q^n} L_{u_1,v} p_{u_2} \left[p_{u_1} p_v - \tfrac{1}{q^2} \left(\sum_{i=1}^{q} r_i \right)^2 \right] \\ &= \sum_{u_1,v=1}^{q^{k+1}} L_{u_1,v} \left[p_{u_1} p_v - \tfrac{1}{q^2} \left(\sum_{i=1}^{q} r_i \right)^2 \right] \sum_{u_2=q^{k+1}+1}^{q^n} p_{u_2} \\ &= (1 - q^k \sum_{i=1}^{q} r_i) \sum_{u,v=1}^{q^{k+1}} L_{u,v} \left[p_u p_v - \tfrac{1}{q^2} \left(\sum_{i=1}^{q} r_i \right)^2 \right]. \end{aligned}$$

We set $A = \sum_{u,v=1}^{q^{k+1}} L_{u,v} \left[p_u p_v - \frac{1}{q^2} \left(\sum_{i=1}^{q} r_i \right)^2 \right]$ and separate the different subtrees with roots in level $n-k-1$ in which u and v can occur. We get

$$A = \sum_{s,t=1}^{q} \sum_{u=(s-1)q^k+1}^{sq^k} \sum_{v=(t-1)q^k+1}^{tq^k} L_{u,v} \left[r_s r_t - \frac{1}{q^2} \left(\sum_{i=1}^{q} r_i \right)^2 \right].$$

Since it holds for $s,t \in [q]$, $u \in [(s-1)q^k+1, sq^k]$ and $v \in [(t-1)q^k+1, tq^k]$ that

$$L_{u,v} = \begin{cases} n-k & \text{if } s \neq t \\ n-k+\mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u,v) & \text{if } s = t, \end{cases}$$

the above equation becomes

$$\begin{aligned} A &= (n-k) \left[\sum_{s,t=1}^{q} q^{2k} r_s r_t - q^{2k} \left(\sum_{i=1}^{q} r_i \right)^2 \right] \\ &\quad + \sum_{s=1}^{q} \sum_{u,v=1}^{q^k} \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u,v) \left[r_s^2 - \tfrac{1}{q^2} \left(\sum_{i=1}^{q} r_i \right)^2 \right] \\ &= \tfrac{1}{2q} \sum_{i,j=1}^{q} (r_i - r_j)^2 \sum_{u,v=1}^{q^k} \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u,v). \end{aligned}$$

The second equation follows on the one hand from $\sum_{s,t=1}^{q} r_s r_t = \left(\sum_{i=1}^{q} r_i \right)^2$. From this follows that the first summand is 0. On the other hand

$$\sum_{s=1}^{q} r_s^2 - \frac{1}{q} \left(\sum_{i=1}^{q} r_i \right)^2 = \frac{1}{2q} \sum_{i,j=1}^{q} (r_i - r_j)^2.$$

By applying Corollary 3.3 we obtain

$$\begin{aligned} \sum_{u,v=1}^{q^k} \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u,v) &= q^k \sum_{l=1}^{k} l |\mathcal{R}^{1,q}_{\mathcal{C}_{q^k}}(q^k, l, 1)| \\ &= q^k \left[\sum_{l=1}^{k-1} l q^{k-l}(q-1) + kq \right] \\ &= q^k \left[q^k (q-1) \sum_{l=1}^{k} l q^{-l} + k \right] \\ &= q^k \left[q^k (q-1) \tfrac{q(q^k-1)-k(q-1)}{q^k (q-1)^2} + k \right] \\ &= \tfrac{q}{q-1} q^k (q^k - 1). \end{aligned}$$

Putting all this together we get

$$R_3 = \frac{1}{q-1} q^k (q^k - 1)(1 - q^k \sum_{i=1}^{q} r_i) \sum_{i,j=1}^{q} (r_i - r_j)^2 \geq 0.$$

This equals 0 if and only if either $k = 0$ or $r_i = r_j$ for all $i, j \in [q]$ or $\sum_{i=1}^{q} r_i = q^{-k}$. The last condition is equivalent to $p_i = 0$ for all $i \in [q^{k+1}+1, q^n]$.

We now turn to R_1. With the same notation as before we have

$$\begin{aligned} R_1 &= \sum_{u_1,u_2,v=1}^{q^{k+1}} L_{u_1u_2,v} \left[p_{u_1} p_{u_2} p_v - \tfrac{1}{q^3} \left(\sum_{i=1}^{q} r_i \right)^3 \right] \\ &= \sum_{s_1,s_2,t=1}^{q} \sum_{r=1}^{2} \sum_{u_r=(s_r-1)q^k+1}^{s_r q^k} \sum_{v=(t-1)q^k+1}^{tq^k} L_{u_1u_2,v} \left[r_{s_1} r_{s_2} r_t - \tfrac{1}{q^3} \left(\sum_{i=1}^{q} r_i \right)^3 \right]. \\ &= \sum_{s_1,s_2,t=1}^{q} \left[r_{s_1} r_{s_2} r_t - \tfrac{1}{q^3} \left(\sum_{i=1}^{q} r_i \right)^3 \right] \sum_{r=1}^{2} \sum_{u_r=(s_r-1)q^k+1}^{s_r q^k} \sum_{v=(t-1)q^k+1}^{tq^k} L_{u_1u_2,v}. \end{aligned}$$

For $u_r \in [(s_r - 1)q^k + 1, s_r q^k]$ and $v \in [(t-1)q^k + 1, tq^k]$ it holds that

$$L_{u_1u_2,v} = \begin{cases} n - k & \text{if } s_1 \neq t \text{ and } s_2 \neq t \\ n - k + \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u_1, v) & \text{if } s_1 = t \text{ and } s_2 \neq t \\ n - k + \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u_2, v) & \text{if } s_1 \neq t \text{ and } s_2 = t \\ n - k + \mathcal{L}^{2,q}_{\mathcal{C}_{q^k}}((u_1, u_2), v) & \text{if } s_1 = s_2 = t. \end{cases}$$

If we insert the above equations into R_1, we get

$$\begin{aligned}
R_1 &= (n-k)\left[\sum_{s_1,s_2,t=1}^{q} q^{3k} r_{s_1} r_{s_2} r_t - q^{3k}\left(\sum_{i=1}^{q} r_i\right)^3\right] \\
&\quad + \sum_{s_1=1}^{q} \sum_{s_2=1, s_2\neq s_1}^{q} q^k \sum_{u_1,v=1}^{q^k} \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u_1,v)\left[r_{s_1}^2 r_{s_2} - \tfrac{1}{q^3}\left(\sum_{i=1}^{q} r_i\right)^3\right] \\
&\quad + \sum_{s_2=1}^{q} \sum_{s_1=1, s_1\neq s_2}^{q} q^k \sum_{u_2,v=1}^{q^k} \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u_2,v)\left[r_{s_1} r_{s_2}^2 - \tfrac{1}{q^3}\left(\sum_{i=1}^{q} r_i\right)^3\right] \\
&\quad + \sum_{s=1}^{q} \sum_{u_1,u_2,v=1}^{q^k} \mathcal{L}^{2,q}_{\mathcal{C}_{q^k}}((u_1,u_2),v)\left[r_s^3 - \tfrac{1}{q^3}\left(\sum_{i=1}^{q} r_i\right)^3\right] \\
&= 2q^k\left[\sum_{s=1}^{q} \sum_{t=1, t\neq s}^{q} r_s^2 r_t - \tfrac{q-1}{q^2}\left(\sum_{i=1}^{q} r_i\right)^3\right] \sum_{u,v=1}^{q^k} \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u,v) \\
&\quad + \left[\sum_{s=1}^{q} r_s^3 - \tfrac{1}{q^2}\left(\sum_{i=1}^{q} r_i\right)^3\right] \sum_{u_1,u_2,v=1}^{q^k} \mathcal{L}^{2,q}_{\mathcal{C}_{q^k}}((u_1,u_2),v).
\end{aligned}$$

If all r_i's are zero, we obtain $R_1 = 0$. We exclude this case and normalize the probabilities $r_1, ..., r_q$ by setting $\bar{r}_i = r_i / \sum_{j=1}^{q} r_j$ for $i \in [q]$. This yields

$$\begin{aligned}
R_1 &= \left(\sum_{i=1}^{q} r_i\right)^3 \Bigg[\quad 2q^k\left(\sum_s \sum_{t\neq s} \bar{r}_s^2 \bar{r}_t - \tfrac{q-1}{q^2}\right) \sum_{u,v=1}^{q^k} \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u,v) \\
&\qquad\qquad + \left(\sum_s \bar{r}_s^3 - \tfrac{1}{q^2}\right) \sum_{u_1,u_2,v=1}^{q^k} \mathcal{L}^{2,q}_{\mathcal{C}_{q^k}}((u_1,u_2),v)\Bigg].
\end{aligned}$$

We have already seen during the calculations of R_3 that

$$\sum_{u,v=1}^{q^k} \mathcal{L}^{1,q}_{\mathcal{C}_{q^k}}(u,v) = \frac{q}{q-1} q^k (q^k - 1).$$

By applying Corollary 3.3 we further get that

$$\begin{aligned}
&\sum_{u_1,u_2,v=1}^{q^k} \mathcal{L}^{2,q}_{\mathcal{C}_{q^k}}((u_1,u_2),v) = q^k \sum_{l=1}^{k} l|\mathcal{R}^{2,q}_{\mathcal{C}_{q^k}}(q^k,l,1)| \\
= \;& q^k \left[\sum_{l=1}^{k-1} lq^{2k}\left(2q^{-l}(q-1)(1-q^{-l+1}) + q^{-2l}(q-1)^2\right)\right] \\
& + q^k k\left(2q(q^k-q)+q^2\right) \\
= \;& q^k \left[\sum_{l=1}^{k} lq^{2k}\left(2q^{-l}(q-1)(1-q^{-l+1}) + q^{-2l}(q-1)^2\right)\right] \\
& + q^k k(2q^k-1) \\
= \;& (q-1)q^{3k}\left[2\sum_{l=1}^{k} lq^{-l} - (q+1)\sum_{l=1}^{k} lq^{-2l}\right] \\
& + kq^k(2q^k-1) \\
= \;& (q-1)q^{3k}\left[2\tfrac{q(q^k-1)-k(q-1)}{q^k(q-1)^2} - (q+1)\tfrac{q^2(q^{2k}-1)-k(q^2-1)}{q^{2k}(q^2-1)^2}\right] \\
& + kq^k(2q^k-1) \\
= \;& 2\tfrac{q}{q-1}q^{2k}(q^k-1) - \tfrac{q^2}{q^2-1}q^k(q^{2k}-1) \\
= \;& \tfrac{q}{q-1}q^k(q^k-1)\tfrac{(q+2)q^k-q}{q+1}.
\end{aligned}$$

Applying this result we obtain

$$\begin{aligned}
R_1 \;=\; & \left(\sum_{i=1}^{q} r_i\right)^3 \tfrac{q}{q-1}q^k(q^k-1)\left[2q^k\left(\sum_s \bar{r}_s^2 - \sum_s \bar{r}_s^3 - \tfrac{q-1}{q^2}\right)\right. \\
& \left. + \tfrac{(q+2)q^k-q}{q+1}\left(\sum_s \bar{r}_s^3 - \tfrac{1}{q^2}\right)\right] \\
=\; & -\left(\sum_{i=1}^{q} r_i\right)^3 \tfrac{q}{q-1}q^k(q^k-1)\left[\tfrac{q}{q+1}(q^k+1)\sum_s \bar{r}_s^3 - 2q^k\sum_s \bar{r}_s^2\right. \\
& \left. + \tfrac{(2q+1)q^k-1}{q(q+1)}\right].
\end{aligned}$$

It remains to show that

$$\frac{q}{q+1}(q^k+1)\sum_s \bar{r}_s^3 - 2q^k \sum_s \bar{r}_s^2 + \frac{(2q+1)q^k-1}{q(q+1)} \leq 0.$$

The left hand side obviously equals 0 if $\bar{r}_1 = ... = \bar{r}_q = 1/q$, i.e. $r_1 = ... = r_q$. Let us define $f : \Delta_{q-1} \to \mathbb{R}$ by

$$f(x_1,...,x_{q-1}) = a_1 \left[\sum_{s=1}^{q-1} x_s^3 + \left(1 - \sum_{s=1}^{q-1} x_s\right)^3\right] - a_2 \left[\sum_{s=1}^{q-1} x_s^2 + \left(1 - \sum_{s=1}^{q-1} x_s\right)^2\right],$$

where $a_1 = q(q^k+1)/(q+1)$ and $a_2 = 2q^k$. We will show that $(1/q,...,1/q)$ is the only extremal point of f in Γ_q and that it is a local maximum. The first partial derivative for $j \in [q-1]$ is

$$\begin{aligned}\tfrac{\delta}{\delta x_j} f(x_1,...,x_{q-1}) &= 3a_1\left(x_j^2 - (1-\sum_{i=1}^{q-1} x_i)^2\right) - 2a_2\left(x_j - (1-\sum_{i=1}^{q-1} x_i)\right) \\ &= \left(x_j - (1-\sum_{i=1}^{q-1} x_i)\right)\left[3a_1\left(x_j + 1 - \sum_{i=1}^{q-1} x_i\right) - 2a_2\right].\end{aligned}$$

It follows that the gradient $\nabla f = \mathbf{0}$ if and only if either $x_j = 1 - \sum_{i=1}^{q-1} x_i$ for all $j \in [q-1]$, which yields $x_1 = ... = x_{q-1} = 1/q$, or $3a_1(x_j + 1 - \sum_{i=1}^{q-1} x_i) - 2a_2 = 0$ for all $j \in [q-1]$. Since

$$3a_1(1 - \sum_{i=1, i\neq j}^{q-1} x_i) - 2a_2 \leq 3\frac{q}{q+1}(q^k+1) - 4q^k < -q^k + 3 \leq 0,$$

the latter is impossible. We conclude that the only extremal point of f is $(1/q,..,1/q)$. Further, the second partial derivatives are

$$\frac{\delta^2}{\delta x_k \delta x_j} f(x_1,...,x_{q-1}) = \begin{cases} 6a_1(1 - \sum_{i=1}^{q-1} x_i) - 2a_2 & \text{if } k \neq j \\ 6a_1(1 - \sum_{i=1, i\neq j}^{q-1} x_i) - 4a_2 & \text{if } k = j \end{cases}$$

such that

$$\frac{\delta^2}{\delta x_k \delta x_j} f\left(\frac{1}{q},..,\frac{1}{q}\right) = \begin{cases} \frac{6a_1}{q} - 2a_2 & \text{if } k \neq j \\ \frac{12a_1}{q} - 4a_2 & \text{if } k = j. \end{cases}$$

Since $(6a_1/q) - 2a_2 = [6(q^k+1)/(q+1)] - 4q^k \leq -2(q^k-1) < 0$, we see that $(1/q,..,1/q)$ is a global maximum. With this we obtain that $R_1 \geq 0$, with

equality if and only if either $k = 0$ or $r_i = r_j$ for all $i, j \in [q]$. Remember that $R_3 \geq 0$. It equals zero if and only if either $k = 0$ or $r_i = r_j$ for all $i, j \in [q]$ or $p_i = 0$ for $i \in [q^{k+1} + 1, q^n]$. Further, $\mathcal{L}^{2,q}_{\mathcal{C}_{q^n}}(P, P) - \mathcal{L}^{2,q}_{\mathcal{C}_{q^n}}(\tilde{P}, \tilde{P}) = R_1 + R_3$. It follows that this difference is not negative. Moreover, it equals 0 if and only if either $k = 0$ or $r_i = r_j$ for all $i, j \in [q]$. This concludes the proof.

□

By applying Lemma 4.3 in the same way as Lemma 2.1 in Subsection 2.1 we obtain

Corollary 4.4 *Let $n \in \mathbb{N}$ and $q \in \mathbb{N}_{\geq 2}$. Further, let $\mathcal{C} = \mathcal{C}_{q^n}$ and $T = T_{\mathcal{C}}$. Then it holds for all probability distributions P on $[q^n]$ that*

$$\mathcal{L}^{2,q}_{\mathcal{C}}(P, P) \geq \mathcal{L}^{2,q}_{\mathcal{C}}\left((\frac{1}{q^n}, ..., \frac{1}{q^n}), (\frac{1}{q^n}, ..., \frac{1}{q^n})\right).$$

The inequality holds with equality if and only if $P(T_x) = q^{-\|x\|}$ for all inner nodes $x \in \mathring{\mathcal{N}}(T)$.

Before we come to Lemma 4.5, we provide a short excurs on *δ-typical sequences*. These are defined e.g. in [8] Definition 2.8 (p. 33). We will change some of the notation of this definition in order to harmonize it with the notation used in this thesis and related papers.

"For any distribution P on $\mathcal{U}$, a sequence $u^n \in \mathcal{U}^n$ is called *P-typical* with constant δ if

$$\left|\frac{1}{n} < u^n | a > -p_a\right| \leq \delta \tag{4.3}$$

for every $a \in \mathcal{U}$ and, in addition, no $a \in \mathcal{U}$ with $p_a = 0$ occurs in u^n. The set of such sequences will be denoted by $\mathcal{T}^n_{P,\delta}$."

Here, the value of $< u^n | a >$ is the number of appearances of a as a component of u^n. In words, a sequence $u^n \in \mathcal{U}^n$ is called *P-typical* with constant δ if for all $a \in \mathcal{U}$ the difference between the relative frequency of a in u^n and the actual probability of a with respect to P is at most δ.

Lemma 2.12 in [8] and its subsequent remark state that

$$P^n(\mathcal{T}^n_{P,\delta}) \geq 1 - \frac{|\mathcal{U}|}{4n\delta^2} \tag{4.4}$$

Further, it follows from Equation (4.3) for all $u^n \in \mathcal{T}^n_{P,\delta}$ that

$$P^n_{u^n} = \prod_{a \in \mathcal{U}} p_a^{< u^n | a >} \leq \prod_{a \in supp(P)} p_a^{n(p_a - \delta)} = 2^{-n\left(H(P) + \delta \sum\limits_{a \in supp(P)} \log p_a\right)}. \tag{4.5}$$

Here, $H(P) = -\sum_{a \in supp(P)} p_a \log p_a$ is Shannon's classical entropy. In the following we use $M_P = -\sum_{a \in supp(P)} \log p_a$. It holds that $0 \leq M_P < \infty$ with equality on the left hand side if and only if $supp(P) = 1$. We exclude this case in our further analysis. It follows that for all $\epsilon > 0$ exists $\delta > 0$ such that on the one hand it holds that

$$P^n((\mathcal{T}^n_{P,\delta})^c) \leq \frac{|\mathcal{U}| M_P}{4n\epsilon^2}. \tag{4.6}$$

On the other hand it holds for all $u^n \in \mathcal{T}^n_{P,\delta}$ that

$$P^n_{u^n} \leq 2^{-n(H(P)-\epsilon)}. \tag{4.7}$$

To see this choose $\delta = \epsilon / M_P$ and apply Equations (4.4) and (4.5). Things are now settled to prove

Lemma 4.5 *Let P be probability distribution on $\mathcal{U}$ with $|supp(P)| > 1$. For all $\epsilon > 0$ and all q-ary prefix codes $\mathcal{C}$ over $\mathcal{U}$ there exist sequences $\alpha_n(\epsilon) = \alpha_n \to 0$ and $K_n(\epsilon) = K_n \to \infty$ such that*

$$\mathcal{L}^{2,q}_{\mathcal{C}^n}(P^n, P^n) \geq (1-\alpha_n)^3 \mathcal{L}^{2,q}_{\mathcal{C}_{q^{K_n}}} \left((\frac{1}{q^{K_n}}, ..., \frac{1}{q^{K_n}}), (\frac{1}{q^{K_n}}, ..., \frac{1}{q^{K_n}}) \right)$$

holds for all sufficiently large n.

Proof:

The proof of this theorem follows the same guidelines as the proof of Lemma 3 in [5]. However, we changed some of its steps in order to obtain a more explanatory proof.

We begin the proof without explicitly specifying K_n and α_n. This will be done later. We partition $\mathcal{U}^n$ according to the given code $\mathcal{C}^n$ into $\mathcal{U}^n_1 = \{u^n \in \mathcal{U}^n \ : \ \|c_{u^n}\| \leq K_n\}$ and $\mathcal{U}^n_2 = \mathcal{U}^n \backslash \mathcal{U}^n_1$. Since $\mathcal{C}^n$ is a q-ary prefix code, we have that

$$|\mathcal{U}^n_1| \leq q^{K_n}. \tag{4.8}$$

For $\epsilon > 0$ we choose $\delta = \epsilon / M_P$ and obtain

$$\begin{aligned} P^n(\mathcal{U}^n_1) &= P^n(\mathcal{U}^n_1 \cap \mathcal{T}^n_{P,\delta}) + P^n(\mathcal{U}^n_1 \cap (\mathcal{T}^n_{P,\delta})^c) \\ &\leq |\ \mathcal{U}^n_1 \cap \mathcal{T}^n_{P,\delta}| 2^{-n(H(P)-\epsilon)} + P^n((\mathcal{T}^n_{P,\delta})^c) \\ &\leq q^{K_n} 2^{-n(H(P)-\epsilon)} + \tfrac{|\mathcal{U}| M_P}{4n\epsilon^2}. \end{aligned}$$

The first inequality follows by Equation (4.7). Equations (4.6) and (4.8) yield the second inequality.

We now set $K_n = \left\lfloor \frac{n(H(P)-2\epsilon)}{\log q} \right\rfloor$ as well as $\alpha_n = 2^{-n\epsilon} + \frac{|\mathcal{U}|M_P}{4n\epsilon^2}$ and obtain

$$P^n(\mathcal{U}_1^n) \leq \alpha_n$$

and thus

$$P^n(\mathcal{U}_2^n) \geq 1 - \alpha_n. \tag{4.9}$$

We will now construct a new source code by cutting all codewords in $\mathcal{U}_2^n$ back to length K_n. Formally, we define the new source $\tilde{\mathcal{U}} = \tilde{\mathcal{U}}_1 \cup \tilde{\mathcal{U}}_2$, where $\tilde{\mathcal{U}}_1 = \mathcal{U}_1^n$ and $\tilde{\mathcal{U}}_2$ is defined as follows. Let $\cong$ be an equivalence relation on $\mathcal{U}_2^n$ with $u^n \cong v^n :\Leftrightarrow c_{u^n}^{K_n} = c_{v^n}^{K_n}$ and let $\mathcal{E}_1, ..., \mathcal{E}_m$ be the equivalence classes. Further, we associate with every equivalence class $\mathcal{E}_i$ the object e_i and define $\tilde{\mathcal{U}}_2 = \{e_1, ..., e_m\}$. Moreover, we define a probability distribution $\tilde{P}$ on $\tilde{\mathcal{U}}$ by $\tilde{P}(u^n) = P(u^n)$ for all $u^n \in \tilde{\mathcal{U}}_1$ and $\tilde{P}(e_k) = \sum_{u^n \in \mathcal{E}_k} P(u^n)$ for $k \in [m]$. Finally, we obtain a new code $\tilde{\mathcal{C}} : \tilde{\mathcal{U}} \to \mathcal{Q}^*$ by $\tilde{c}_{u^n} = c_{u^n}$ if $u^n \in \tilde{\mathcal{U}}_1$ and $\tilde{c}_{e_k}$ will be the common prefix of length K_n of the objects in $\mathcal{E}_k$. This construction step is visualized in Figure 4.1. It follows that

$$\mathcal{L}_{\mathcal{C}^n}^{2,q}(P^n, P^n) \geq \mathcal{L}_{\tilde{\mathcal{C}}}^{2,q}(\tilde{P}, \tilde{P}). \tag{4.10}$$

The next step is to focus only on the $\tilde{\mathcal{U}}_2$-part of $\tilde{\mathcal{U}}$. Again we operate without increasing the symmetric 2-identification running time since

$$\begin{aligned}
\mathcal{L}_{\tilde{\mathcal{C}}}^{2,q}(\tilde{P}, \tilde{P}) &= \sum_{\tilde{u}_1, \tilde{u}_2, \tilde{v} \in \tilde{\mathcal{U}}} \tilde{P}(\tilde{u}_1)\tilde{P}(\tilde{u}_2)\tilde{P}(\tilde{v})\mathcal{L}_{\tilde{\mathcal{C}}}^{2,q}((\tilde{u}_1, \tilde{u}_2), \tilde{v}) \\
&\geq \sum_{\tilde{u}_1, \tilde{u}_2, \tilde{v} \in \tilde{\mathcal{U}}_2} \tilde{P}(\tilde{u}_1)\tilde{P}(\tilde{u}_2)\tilde{P}(\tilde{v})\mathcal{L}_{\tilde{\mathcal{C}}}^{2,q}((\tilde{u}_1, \tilde{u}_2), \tilde{v}) \\
&= \sum_{i_1, i_2, j=1}^{m} \tilde{P}(e_{i_1})\tilde{P}(e_{i_2})\tilde{P}(e_j)\mathcal{L}_{\tilde{\mathcal{C}}}^{2,q}((e_{i_1}, e_{i_2}), e_j) \\
&= \left(\sum_{k=1}^{m} \tilde{P}(e_k)\right)^3 \sum_{i_1, i_2, j=1}^{m} \tilde{P}_2(e_{i_1})\tilde{P}_2(e_{i_2})\tilde{P}_2(e_j)\mathcal{L}_{\tilde{\mathcal{C}}_2}^{2,q}((e_{i_1}, e_{i_2}), e_j).
\end{aligned}$$

Here, $\tilde{P}_2$ is a probability distribution on $\tilde{\mathcal{U}}_2$ defined by $\tilde{P}_2(e_j) = \tilde{P}(e_j) / \sum_{k=1}^{m} \tilde{P}(e_k)$ for $j \in [m]$. Further, $\tilde{\mathcal{C}}_2$ is the restriction of $\tilde{\mathcal{C}}$ to $\tilde{\mathcal{U}}_2$.

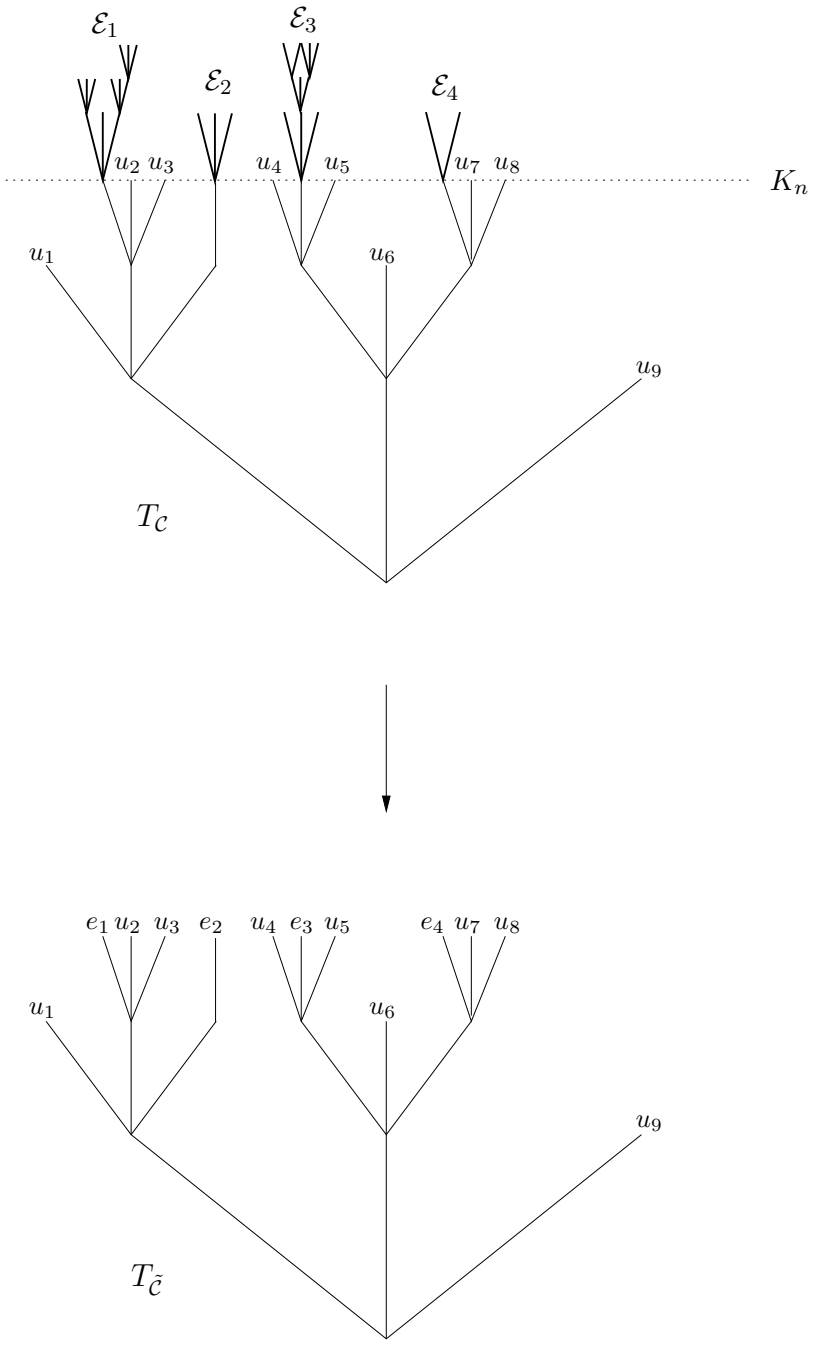

Figure 4.1: The cutting of $T_{\mathcal{C}}$ at depth K_n yields $T_{\tilde{\mathcal{C}}}$ with $\tilde{\mathcal{U}}_1 = \{u_1, u_2, ..., u_9\}$ and $\tilde{\mathcal{U}}_2 = \{e_1, e_2, e_3, e_4\}$.

Since

$$\sum_{k=1}^{m} \tilde{P}(e_k) = \sum_{k=1}^{m} \sum_{u^n \in \mathcal{E}_k} P^n(u^n) = P^n(\mathcal{U}_2^n),$$

we obtain by Equation (4.9) that

$$\mathcal{L}_{\tilde{\mathcal{C}}}^{2,q}(\tilde{P}, \tilde{P}) \geq (1-\alpha_n)^3 \mathcal{L}_{\tilde{\mathcal{C}}_2}^{2,q}(\tilde{P}_2, \tilde{P}_2). \tag{4.11}$$

Although $\tilde{\mathcal{C}}_2$ is a block code with codewords of length K_n it may be - and maybe by far - not saturated. To achieve this property we extend $\tilde{\mathcal{U}}_2$ to a set of cardinality q^{K_n}, assign zero probabilities to the additional elements and use for them codewords from $\mathcal{Q}^{K_n} \backslash \tilde{\mathcal{C}}_2$. We now obey the conditions of Corollary 4.4 by which we obtain

$$\mathcal{L}_{\tilde{\mathcal{C}}_2}^{2,q}(\tilde{P}_2, \tilde{P}_2) \geq \mathcal{L}_{\mathcal{C}_{q^{K_n}}}^{2,q}((\frac{1}{q^{K_n}}, ..., \frac{1}{q^{K_n}}), (\frac{1}{q^{K_n}}, ..., \frac{1}{q^{K_n}})). \tag{4.12}$$

The inequalities (4.10), (4.11) and (4.12) finally yield the statement of the lemma.

□

By applying Theorem 3.4 and Lemma 4.5 to Corollary 4.2 we obtain

Corollary 4.6 *Let $\mathcal{U}$ be a finite set, $q \in \mathbb{N}_{\geq 2}$, P be a probability distribution on $\mathcal{U}$ with $|supp(P)| > 1$ and $\mathcal{C}$ be a q-ary prefix code. It then holds that*

$$\mathcal{L}_{\mathcal{C}}^{2,q}(P,P) \geq (1 - \sum_{u \in \mathcal{U}} p_u^3) \left(2\frac{q}{q-1} - \frac{q^2}{q^2-1} \right) - 2 \left(\frac{1 - \sum\limits_{u \in \mathcal{U}} p_u^3}{1 - \sum\limits_{u \in \mathcal{U}} p_u^2} - 1 \right) \mathcal{L}_{\mathcal{C}}^{1,q}(P,P).$$

4.2 The q-ary Identification Entropy of Second Degree

Since (1-)identification appears negatively signed, we can not immediately apply its lower bound $\mathcal{L}_{\mathcal{C}}^{1,q}(P,P) \geq H_{\mathrm{ID}}^{1,q}(P)$ (see [5]). But we can show that the bound of Corollary 4.6 is attained if P consists only of q-powers and $\mathcal{C}$ is a code with $\|c_u\| = -\log_q p_u$.

Proposition 4.7 *Let P be a probability distribution on $\mathcal{U}$ which only consists of q-powers and $\mathcal{C}$ be a q-ary prefix code, where $\|c_u\| = -\log_q p_u$ for all $u \in \mathcal{U}$. It then holds that*

$$\mathcal{L}_{\mathcal{C}}^{2,q}(P,P) = 2\frac{q}{q-1} \left(1 - \sum_{u \in \mathcal{U}} p_u^2 \right) - \frac{q^2}{q^2-1} \left(1 - \sum_{u \in \mathcal{U}} p_u^3 \right).$$

Proof:

It is an immediate consequence from the condition $\|c_u\| = -\log_q p_u$ for all $u \in \mathcal{U}$ that

$$P(T_x) = q^{-\|x\|} \tag{4.13}$$

holds for all $x \in \mathcal{N}(T)$, where $T = T_{\mathcal{C}}$. We now introduce for all $v \in \mathcal{U}$ and $k = 1, ..., \|c_v\|$ the set

$$\bar{\mathcal{R}}_{\mathcal{C}}^{1,q}(k, v) = \mathcal{R}_{\mathcal{C}}^{1,q}(1, v) \dot{\cup} ... \dot{\cup} \mathcal{R}_{\mathcal{C}}^{1,q}(k-1, v). \tag{4.14}$$

Proceeding as in the proof of Theorem 3.4 we obtain

$$\mathcal{L}_{\mathcal{C}}^{L,q}(P, P) = \sum_{v \in \mathcal{U}} p_v \sum_{k=1}^{\|c_v\|} k \sum_{(u_1,u_2) \in \mathcal{R}_{\mathcal{C}}^{2,q}(k,v)} p_{u_1} p_{u_2}.$$

In the following we use $S_{k,v} = \sum_{(u_1,u_2) \in \mathcal{R}_{\mathcal{C}}^{2,q}(k,v)} p_{u_1} p_{u_2}$. With the notation of Equation (4.14) it holds that

$$S_{k,v} = 2 \sum_{u_1 \in \mathcal{R}_{\mathcal{C}}^{1,q}(k,v)} \sum_{u_2 \in \bar{\mathcal{R}}_{\mathcal{C}}^{1,q}(k,v)} p_{u_1} p_{u_2} + \sum_{u_1,u_2 \in \mathcal{R}_{\mathcal{C}}^{1,q}(k,v)} p_{u_1} p_{u_2}.$$

Here, the equality holds because there exists either one component for which (1-)identification against v takes exactly k timesteps and the other yields a (1-)identification time regarding v of at most $k-1$ or both components have a (1-)identification time regarding v of k.

Case 1: $\mathbf{k = 1, ..., \|c_v\| - 1}$

In this case we have that $\mathcal{R}_{\mathcal{C}}^{1,q}(k, v) = \bar{T}_{c_v^{k-1}} \backslash \bar{T}_{c_v^k}$ and $\bar{\mathcal{R}}_{\mathcal{C}}^{1,q}(k, v) = \mathcal{U} \backslash \bar{T}_{c_v^{k-1}}$. This together with Equation (4.13) yields

$$\sum_{u \in \mathcal{R}_{\mathcal{C}}^{1,q}(k,v)} p_u = P(T_{c_v^{k-1}}) - P(T_{c_v^k}) = q^{-k+1} - q^{-k} = q^{-k}(q-1)$$

and

$$\sum_{u \in \bar{\mathcal{R}}_{\mathcal{C}}^{1,q}(k,v)} p_u = 1 - P(T_{c_v^{k-1}}) = 1 - q^{-k+1}.$$

Thus,

$$\begin{aligned} S_{k,v} &= 2q^{-k}(q-1)(1-q^{-k+1}) + q^{-2k}(q-1)^2 \\ &= (1-q^{-k})^2 - (1-q^{-k+1})^2. \end{aligned}$$

Case 2: $\mathbf{k = \|c_v\|}$

In this case we have that $\mathcal{R}_{\mathcal{C}}^{1,q}(\|c_v\|, v) = \bar{T}_{c_v^{\|c_v\|-1}}$ and $\bar{\mathcal{R}}_{\mathcal{C}}^{1,q}(\|c_v\|, v) = \mathcal{U} \backslash \bar{T}_{c_v^{\|c_v\|-1}}$. Equation (4.13) yields

$$\sum_{u \in \mathcal{R}_{\mathcal{C}}^{1,q}(\|c_v\|,v)} p_u = P(T_{c_v^{\|c_v\|-1}}) = q^{-\|c_v\|+1}$$

and

$$\sum_{u \in \bar{\mathcal{R}}_{\mathcal{C}}^{1,q}(\|c_v\|,v)} p_u = 1 - P(T_{c_v^{\|c_v\|-1}}) = 1 - q^{-\|c_v\|+1}.$$

Thus, we obtain

$$S_{\|c_v\|,v} = 2q^{-\|c_v\|+1}(1 - q^{-\|c_v\|+1}) + q^{-2(\|c_v\|-1)} = 1 - (1 - q^{-\|c_v\|+1})^2.$$

Together, the above two cases yield

$$\begin{aligned}
& \sum_{k=1}^{\|c_v\|} k S_{k,v} \\
= & \sum_{k=1}^{\|c_v\|-1} k \left[(1-q^{-k})^2 - (1-q^{-k+1})^2\right] + \|c_v\| \left[1 - (1-q^{-\|c_v\|+1})^2\right] \\
= & \sum_{k=1}^{\|c_v\|-1} k(1-q^{-k})^2 + \|c_v\| - \sum_{k=1}^{\|c_v\|} k(1-q^{-k+1})^2.
\end{aligned}$$

If we take a look at the first sum plus $\|c_v\|$, we see that

$$\begin{aligned}
\sum_{k=1}^{\|c_v\|-1} k(1-q^{-k})^2 + \|c_v\| &= \sum_{k=1}^{\|c_v\|-1} k(1 - 2q^{-k} + q^{-2k}) + \|c_v\| \\
&= \sum_{k=1}^{\|c_v\|} k - 2 \sum_{k=1}^{\|c_v\|-1} k q^{-k} + \sum_{k=1}^{\|c_v\|-1} k q^{-2k}.
\end{aligned}$$

Further, we obtain

$$\begin{aligned}
\sum_{k=1}^{\|c_v\|} k(1-q^{-k+1})^2 &= \sum_{k=1}^{\|c_v\|} k(1 - 2q^{-k+1} + q^{-2k+2}) \\
&= \sum_{k=1}^{\|c_v\|} k - 2 \sum_{k=1}^{\|c_v\|} k q^{-k+1} + \sum_{k=1}^{\|c_v\|} k q^{-2k+2}.
\end{aligned}$$

Subtracting the second from the first result we get

$$\begin{aligned}
\sum_{k=1}^{\|c_v\|} kS_{k,v} &= 2(q-1)\sum_{k=1}^{\|c_v\|} kq^{-k} - (q^2-1)\sum_{k=1}^{\|c_v\|} kq^{-2k} \\
&\quad +\|c_v\|q^{-\|c_v\|}(2-q^{-\|c_v\|}) \\
&= 2\tfrac{q}{q-1}(1-p_v) - 2\|c_v\|p_v - \tfrac{q^2}{q^2-1}(1-p_v^2) + \|c_v\|p_v^2 \\
&\quad +\|c_v\|p_v(2-p_v) \\
&= 2\tfrac{q}{q-1}(1-p_v) - \tfrac{q^2}{q^2-1}(1-p_v^2).
\end{aligned}$$

Here, the first equality follows from the previously calculated sums. The second equality holds since by assumption $q^{-\|c_v\|} = p_v$ for all $v \in \mathcal{U}$ and since we have for $j = 1, 2$ that

$$\begin{aligned}
\sum_{k=1}^{\|c_v\|} kq^{-jk} &= \tfrac{1}{(q^j-1)^2}[q^j - (q^j(\|c_v\|+1) - \|c_v\|)q^{-j\|c_v\|}] \\
&= \tfrac{q^l}{(q^l-1)^2}(1-p_v^l) - \tfrac{\|c_v\|}{q^l-1}p_v^l.
\end{aligned}$$

Finally the above calculations yield

$$\begin{aligned}
\mathcal{L}_{\mathcal{C}}^{L,q}(P,P) &= \sum_{v\in\mathcal{U}} p_v \sum_{k=1}^{\|c_v\|} kS_{k,v} \\
&= 2\tfrac{q}{q-1}\left(1 - \sum_{v\in\mathcal{U}} p_v^2\right) - \tfrac{q^2}{q^2-1}\left(1 - \sum_{v\in\mathcal{U}} p_v^3\right).
\end{aligned}$$

□

This result encourages us in the believe that the right side of the equation in Proposition 4.7 is in general a lower bound for 2-identification. As we will see soon it obeys some fundamental properties for entropy functions. Therefore, we define $H_{\mathrm{ID}}^{2,q} : \Gamma_N \to \mathbb{R}$ by

$$H_{\mathrm{ID}}^{2,q}(P) = 2\frac{q}{q-1}\left(1 - \sum_{u\in\mathcal{U}} p_u^2\right) - \frac{q^2}{q^2-1}\left(1 - \sum_{u\in\mathcal{U}} p_u^3\right). \tag{4.15}$$

We call it the *q-ary identification-entropy of second degree*. Its role as a lower bound for 2-identification is expressed in

Theorem 4.8 *Let $\mathcal{U}$ be a finite set and $q \in \mathbb{N}_{\geq 2}$. It holds for all probability distributions P on $\mathcal{U}$ and all q-ary prefix codes $\mathcal{C}$ that*

$$\mathcal{L}_{\mathcal{C}}^{2,q}(P,P) \geq H_{ID}^{2,q}(P),$$

where equality is attained if and only if P consists only of q-powers, and $\mathcal{C}$ is a prefix code, with $\|c_u\| = -\log_q p_u$ for all $u \in \mathcal{U}$.

Before we prove Theorem 4.8, we will first analyze the functional properties of $H_{\mathrm{ID}}^{2,q}$. A list of desiderata for entropy functions can be found in [1], pp. 50. We now show that entropy function obeys important ones of them.

Theorem 4.9 *The following properties hold for $H_{ID}^{2,q}(P)$.*

1. *Symmetry:*
$$H_{ID}^{2,q}(p_1,...,p_N) = H_{ID}^{2,q}(p_{\pi(1)},...,p_{\pi(N)}), \tag{4.16}$$
where π is a permutation on $[N]$.

2. *Expansibility:*
$$H_{ID}^{2,q}(p_1,...,p_N) = H_{ID}^{2,q}(p_1,...,p_N,0). \tag{4.17}$$

3. *Decisiveness:*
$$H_{ID}^{2,q}(1,0,...,0) = 0.$$

4. *Normalization:*
$$H_{ID}^{2,q}\left(\frac{1}{q},...,\frac{1}{q}\right) = 1. \tag{4.18}$$

5. *Bounds:*
$$H_{ID}^{2,q}(1,0,...,0) \leq H_{ID}^{2,q}(P) \leq H_{ID}^{2,q}\left(\frac{1}{N},...,\frac{1}{N}\right). \tag{4.19}$$

6. *Grouping Behavior: For $m \leq N$ let*

   a) $\mathcal{U}_1, \mathcal{U}_2, ..., \mathcal{U}_m$ be a partition of $\mathcal{U}$ of non-empty sets

   b) $Q = (Q_1,...,Q_m)$ be the probability distribution on $[m]$ defined by $Q_i = \sum_{u \in \mathcal{U}_i} p_u$

   c) P_i is the probability distribution on $\mathcal{U}_i$ defined by $p_{i,u} = p_u/Q_i$ for all $i \in [m]$ and $u \in \mathcal{U}_i$.

   It then holds that

$$H_{ID}^{2,q}(P) = H_{ID}^{2,q}(Q) + \sum_{i=1}^{m} \left[2Q_i^2(1-Q_i)H_{ID}^{1,q}(P_i) + Q_i^3 H_{ID}^{2,q}(P_i)\right]. \tag{4.20}$$

Proof:

Symmetry, expansibility and decisiveness follow directly from the definition of $H_{\mathrm{ID}}^{2,q}$. Further, the normalization property follows from

$$H_{\mathrm{ID}}^{2,q}\left(\frac{1}{q},...,\frac{1}{q}\right) = 2\frac{q}{q-1}\left(1-\frac{1}{q}\right) - \frac{q^2}{q^2-1}\left(1-\frac{1}{q^2}\right) = 1.$$

Bounds:
Let $f(p_1,...,p_{N-1}) = H_{\mathrm{ID}}^{2,q}(p_1,...,p_{N-1},1-\sum_{i=1}^{N-1}p_i)$. We will show that the gradient $\nabla f(p_1,...,p_{N-1}) = \mathbf{0}$ if and only if $(p_1,...,p_{N-1}) = (1/N,...,1/N)$. For that we set $p_N = 1-\sum_{i=1}^{N-1}p_i$ and obtain that it holds for all $j \in [N-1]$ that

$$\frac{\delta}{\delta p_j} f(p_1,...,p_{N-1}) = -4\frac{q}{q-1}(p_j - p_N) + 3\frac{q^2}{q^2-1}(p_j^2 - p_N^2).$$

It follows immediately that $\nabla f(1/N,...,1/N) = \mathbf{0}$.

Assume now that for any $P' \neq (1/N,...,1/N)$ it holds that $\nabla f(P') = \mathbf{0}$. It follows that there exists $j \in [N-1]$ such that $p_j \neq p_N$. If we now take a look at $\frac{\delta}{\delta p_j} f(P')$, we see that

$$\frac{\delta}{\delta p_j} f(P') = 0$$

$$\Leftrightarrow \quad 3\tfrac{q}{q+1}(p_j + p_N) = 4.$$

This is a contradiction because $\frac{q}{q+1}(p_j + p_N)$ is clearly smaller than 1.

In order to ensure that $(1/N,...,1/N)$ is indeed a maximum we show that the Hessian is negative definite. In fact, we will obtain a stronger result namely that all second derivatives $\frac{\delta^2}{\delta p_k \delta p_j} f(1/N,...,1/N)$ are strictly negative.

$$\frac{\delta^2}{\delta p_k \delta p_j} f\left(\frac{1}{N},...,\frac{1}{N}\right) = \begin{cases} 4\frac{q}{q-1}(\frac{3q}{N(q+1)} - 2) & \text{if } k = j \\ 2\frac{q}{q-1}(\frac{3q}{N(q+1)} - 2) & \text{if } k \neq j. \end{cases}$$

From $q \geq 2$ now follows that $\frac{3q}{N(q+1)} - 2 < 0$ if $N \geq 2$. And for $N = 1$ we are in the trivial case, where $H_{\mathrm{ID}}^{2,q}(1) = 0$.

Grouping Behavior:
We use

$$S_i = 2Q_i^2(1-Q_i)H_{\mathrm{ID}}^{1,q}(P_i) + Q_i^3 H_{\mathrm{ID}}^{2,q}(P_i),$$

for all $i \in [m]$ and observe that

$$\begin{aligned} S_i &= 2Q_i^2(1-Q_i)\tfrac{q}{q-1}(1-\tfrac{1}{Q_i^2}\sum_{u\in\mathcal{U}_i} p_u^2) \\ &\quad +Q_i^3\left[2\tfrac{q}{q-1}(1-\tfrac{1}{Q_i^2}\sum_{u\in\mathcal{U}_i} p_u^2) - \tfrac{q^2}{q^2-1}(1-\tfrac{1}{Q_i^3}\sum_{u\in\mathcal{U}_i} p_u^3)\right] \\ &= 2\tfrac{q}{q-1}(Q_i^2 - \sum_{u\in\mathcal{U}_i} p_u^2) - \tfrac{q^2}{q^2-1}(Q_i^3 - \sum_{u\in\mathcal{U}_i} p_u^3). \end{aligned}$$

By summing the S_i's up we obtain

$$\sum_{i=1}^m S_i = 2\frac{q}{q-1}(\sum_{i=1}^m Q_i^2 - \sum_{u\in\mathcal{U}} p_u^2) - \frac{q^2}{q^2-1}(\sum_{i=1}^m Q_i^3 - \sum_{u\in\mathcal{U}} p_u^3)$$

and thus

$$\begin{aligned} H_{\mathrm{ID}}^{2,q}(Q) + \sum_{i=1}^m S_i &= 2\tfrac{q}{q-1}(1-\sum_{i=1}^m Q_i^2) - \tfrac{q^2}{q^2-1}(1-\sum_{i=1}^m Q_i^3) \\ &\quad +2\tfrac{q}{q-1}(\sum_{i=1}^m Q_i^2 - \sum_{u\in\mathcal{U}} p_u^2) - \tfrac{q^2}{q^2-1}(\sum_{i=1}^m Q_i^3 - \sum_{u\in\mathcal{U}} p_u^3) \\ &= 2\tfrac{q}{q-1}(1-\sum_{u\in\mathcal{U}} p_u^2) - \tfrac{q^2}{q^2-1}(1-\sum_{u\in\mathcal{U}} p_u^3) \\ &= H_{\mathrm{ID}}^{2,q}(P). \end{aligned}$$

□

In order to prove Theorem 4.8 we need a decomposition formula for the 2-identification running time. It turns out that the decomposition of the 2-identification running time behaves mainly in the same way as the grouping behavior of the q-ary identification entropy of second degree. We prove this formula in its general form since we will also need this lemma in the next section.

Lemma 4.10 *For all $i \in \mathcal{Q}$ let*

1. $\mathcal{U}_i = \{u \in \mathcal{U} \ : \ c_{u,1} = i\}$
2. $Q_i = \sum_{u \in \mathcal{U}_i} p_u$
3. *P_i be a probability distribution on $\mathcal{U}_i$ defined by $p_{i,u} = \frac{p_u}{Q_i}$ for all $u \in \mathcal{U}_i$*
4. *$\mathcal{C}^{(i)} : \mathcal{U}_i \to \mathcal{Q}^*$ be the code on $\mathcal{U}_i$ defined by $c_u^{(i)} = c_{u,2}c_{u,3}...c_{u,\|c_u\|}$ for all $u \in \mathcal{U}_i$.*

Then it holds that

$$\mathcal{L}_{\mathcal{C}}^{L,q}(P,P) = 1 + \sum_{i \in \mathcal{Q}} \sum_{l=1}^{L} \binom{L}{l} Q_i^{l+1}(1-Q_i)^{L-l} \mathcal{L}_{\mathcal{C}^{(i)}}^{l,q}(P_i,P_i).$$

For $L = 2$ this becomes

$$\mathcal{L}_{\mathcal{C}}^{2,q}(P,P) = 1 + \sum_{i \in \mathcal{Q}} \left[2Q_i^2(1-Q_i)\mathcal{L}_{\mathcal{C}^{(i)}}^{1,q}(P_i,P_i) + Q_i^3 \mathcal{L}_{\mathcal{C}^{(i)}}^{2,q}(P_i,P_i)\right].$$

Proof:

We observe that

$$\begin{aligned}\mathcal{L}_{\mathcal{C}}^{L,q}(P,P) &= \sum_{u^L \in \mathcal{U}^L} \sum_{v \in \mathcal{U}} P_{u^L}^L p_v \mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v) \\ &= \sum_{i \in \mathcal{Q}} \sum_{v \in \mathcal{U}_i} \sum_{u^L \in \mathcal{U}^L} P_{u^L}^L p_v \mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v).\end{aligned}$$

Since $\mathcal{L}_{\mathcal{C}}^{L,q}(u^L,v) = \mathcal{L}_{\mathcal{C}}^{L,q}((u_1,...,u_L),v) = \mathcal{L}_{\mathcal{C}}^{L,q}((u_{\pi(1)},...,u_{\pi(L)}),v)$ for all permutations π on $[L]$, we get for all $i \in \mathcal{Q}$

$$\begin{aligned}&\sum_{v \in \mathcal{U}_i} \sum_{u^L \in \mathcal{U}^L} P_{u^L}^L p_v \mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v) \\ =\ & \sum_{l=0}^{L} \tbinom{L}{l} \sum_{v \in \mathcal{U}_i} \sum_{u_1,...,u_l \in \mathcal{U}_i} \sum_{u_{l+1},...,u_L \in \mathcal{U} \setminus \mathcal{U}_i} P_{u^L}^L p_v \mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v) \\ =\ & \sum_{l=0}^{L} \tbinom{L}{l} (1-Q_i)^{L-l} \sum_{u_1,...,u_l,v \in \mathcal{U}_i} p_{u_1}...p_{u_l}p_v(1 + \mathcal{L}_{\mathcal{C}^{(i)}}^{l,q}((u_1,...,u_l),v)) \\ =\ & Q_i \sum_{l=0}^{L} \tbinom{L}{l} Q_i^l (1-Q_i)^{L-l} + \sum_{l=1}^{L} \tbinom{L}{l} (1-Q_i)^{L-l} Q_i^{l+1} \mathcal{L}_{\mathcal{C}^{(i)}}^{l,q}(P_i,P_i) \\ =\ & Q_i + \sum_{l=1}^{L} \tbinom{L}{l} (1-Q_i)^{L-l} Q_i^{l+1} \mathcal{L}_{\mathcal{C}^{(i)}}^{l,q}(P_i,P_i).\end{aligned}$$

The second equality follows since $\mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v) = 1 + \mathcal{L}_{\mathcal{C}^{(i)}}^{l,q}((u_1, ..., u_l), v)$ holds if $u_1, ..., u_l, v \in \mathcal{U}_i$ and $u_{l+1}, ..., u_L \in \mathcal{U} \backslash \mathcal{U}_i$. Adding this up for $i \in \mathcal{Q}$ we obtain the desired result.

□

As one can see there is a strong relation between the above decomposition formula for 2-identification and the grouping behavior of the identification entropy of second degree. In the following inductive proof of Theorem 4.8 we exploit this relation in order to apply the induction step.

Proof of Theorem 4.8:

For $L = 1$ the statement follows for all $N \in \mathbb{N}$ from Theorem 2 in [3]. As the induction base for N we have to consider all the cases $N = 1, .., q$ and since here $\mathcal{L}_{\mathcal{C}}^{2,q}(P, P) = 1$, we have to show that $H_{\mathrm{ID}}^{2,q}(P) \leq 1$. It follows by the expansibility property (4.17) of the second degree identification entropy function that we only have to consider the case $N = q$. Further, the maximality of the uniform distribution (4.19) and the normalization property (4.18) yield

$$H_{\mathrm{ID}}^{2,q}(p_1, ..., p_q) \leq H_{\mathrm{ID}}^{2,q}\left(\frac{1}{q}, ..., \frac{1}{q}\right) = 1.$$

We set $Q = (Q_0, ..., Q_{q-1})$ and use the same notation as in Lemma 4.10. The inequality of Theorem 4.8 now follows from

$$\begin{aligned} \mathcal{L}_{\mathcal{C}}^{2,q}(P, P) &= 1 + \sum_{i \in \mathcal{Q}} \left[2Q_i^2(1 - Q_i)\mathcal{L}_{\mathcal{C}^{(i)}}^{1,q}(P_i, P_i) + Q_i^3 \mathcal{L}_{\mathcal{C}^{(i)}}^{2,q}(P_i, P_i)\right] \\ &\geq H_{\mathrm{ID}}^{2,q}(Q) + \sum_{i \in \mathcal{Q}} \left[2Q_i^2(1 - Q_i)H_{\mathrm{ID}}^{1,q}(P_i) + Q_i^3 H_{\mathrm{ID}}^{2,q}(P_i)\right] \qquad (4.21) \\ &= H_{\mathrm{ID}}^{2,q}(P). \end{aligned}$$

Here, the equality of the first line follows from Lemma 4.10. The inequality is a consequence of the induction step together with the normalization property (4.18) and the established bounds (4.19) of $H_{\mathrm{ID}}^{2,q}$. Finally, the grouping behavior (4.20) of $H_{\mathrm{ID}}^{2,q}$ yields the second equality.

The fact that this lower bound is attained for every q-ary prefix code $\mathcal{C}$ for which equality (4.13) holds has already been proven by Proposition 4.7. If we instead have that the inequality of Theorem 4.8 holds with equality, then also the inequality of equation (4.21) is in fact an equality and thus

i) $H_{\mathrm{ID}}^{2,q}(Q) = 1$

ii) $H_{\mathrm{ID}}^{1,q}(P_i) = \mathcal{L}_{\mathcal{C}_i}^{1,q}(P_i, P_i)$

iii) $H_{\mathrm{ID}}^{2,q}(P_i) = \mathcal{L}_{\mathcal{C}_i}^{2,q}(P_i, P_i)$.

We have seen in the proof of the bounds of the entropy function that the uniform distribution is the only point where the first derivative of the identification entropy function equals zero and thus $(1/q, ..., 1/q)$ is the only point for which $H_{\mathrm{ID}}^{2,q}(Q) = 1$. Together with i) this means that we get for all $i \in \mathcal{Q}$ that

$$Q_i = \frac{1}{q} \tag{4.22}$$

The crucial part is now ii). For all $i \in \mathcal{Q}$ we obtain from Equation (4.22) and the definitions of P_i and $\mathcal{C}^{(i)}$ (see Lemma 4.10) that for $u \in \mathcal{U}_i$ we have

$$p_u = Q_i p_{i,u} = \frac{p_{i,u}}{q} \tag{4.23}$$

and

$$\|c_u\| = \|c_u^{(i)}\| + 1. \tag{4.24}$$

Moreover, Theorem 1 in [5] stated that for (1-)identification an equality between the running time and identification entropy is only attained if and only if the probability distribution consists only of q-powers and the lengths of the codewords equal the negative logarithm of the probability of their corresponding elements. Thus it follows from ii) that all the $p_{i,u}$'s are q-powers and that $\|c_u^{(i)}\| = -\log_q p_{i,u}$. Together with Equations (4.23) and (4.24) we finally obtain that P consists only of q-powers and that

$$\|c_u\| = -\log_q p_{i,u} + 1 = -\log_q \frac{p_{i,u}}{q} = -\log_q p_u.$$

□

In Theorem 3.4 we have shown for the uniform distribution that if $\mathcal{C}$ is a balanced Huffman code, its symmetric 2-identification running time asymptotically equals

$$K_{2,q} = 2\frac{q}{q-1} - \frac{q^2}{q^2-1}.$$

Since

$$\begin{aligned} H_{\mathrm{ID}}^{2,q}\left(\tfrac{1}{N},...,\tfrac{1}{N}\right) &= 2\tfrac{q}{q-1}\tfrac{N-1}{N} - \tfrac{q^2}{q^2-1}\tfrac{N^2-1}{N^2} \\ &= 2\tfrac{q}{q-1} - \tfrac{q^2}{q^2-1} - 2\tfrac{q}{q-1}\tfrac{1}{N} + \tfrac{q^2}{q^2-1}\tfrac{1}{N^2} \end{aligned}$$

and thus

$$\lim_{N\to\infty} H_{\mathrm{ID}}^{2,q}\left(\frac{1}{N},...,\frac{1}{N}\right) = K_{2,q},$$

we get

Corollary 4.11 *Considering the uniform distribution, balanced Huffman codes are asymptotically optimal for 2-identification.*

4.3 An Upper Bound for Binary Codes

In this subsection we establish an upper bound for $q = 2$. As said in the introduction of this section this is done mainly by the same code construction as in Subsection 2.2. We define $\mathcal{U}_{\max}$, $p_{\max}$ and $P_{\max}$ according to Equations (2.2), (2.3) and (2.4). Further, Equation (2.5) becomes

$$\mathcal{L}_{\mathcal{C}}^{2,2}(P) \leq 1 + 2(1-p_{\max})p_{\max}\mathcal{L}_{\mathcal{C}_{\max}}^{1,2}(P_{\max}) + p_{\max}^2\mathcal{L}_{\mathcal{C}_{\max}}^{2,2}(P_{\max}). \tag{4.25}$$

We prove now by induction over N the following

Theorem 4.12 *It holds for all probability distributions P on $\mathcal{U}$ that the worst-case running time for binary 2-identification can be upper bounded by*

$$\mathcal{L}^{2,2}(P) < \frac{55}{16}.$$

Proof:

W.l.o.g. we assume that $p_1 \geq p_2 \geq ... \geq p_N$. As induction base serve the cases $N = 1, 2$ for which the running time always equals 1.

In order to apply the upper bound for (1-)identification, we use the same code construction as in Theorem 2.3. We partition $\mathcal{U}$ into sets $\mathcal{U}_0$ and $\mathcal{U}_1$, which differ from case to case. We choose t such that $|\frac{1}{2} - \sum_{u=1}^{t} p_u|$ is minimal and set

$$\mathcal{U}_0 = \begin{cases} \{1\} & \text{if } p_1 \geq \frac{1}{2} \\ \{1,2\} & \text{if } p_1 < \frac{1}{2} \text{ and } t = 1 \\ \{1,...,t\} & \text{if } p_1 < \frac{1}{2} \text{ and } t \geq 2. \end{cases}$$

Once we have chosen $\mathcal{U}_0$ and $\mathcal{U}_1 = \mathcal{U} \backslash \mathcal{U}_0$ we inductively construct codes $\mathcal{C}_i$ on $\mathcal{U}_i$. Note that $\mathcal{C}_0 = \emptyset$ if $p_1 \geq 1/2$. From these codes we derive a code $\mathcal{C}$ on $\mathcal{U}$ by prefixing all codewords in $\mathcal{C}_i$ with i.

Case 1: $\mathbf{p_1 \geq \frac{1}{2}}$
For the same reason as in the proof of Theorem 2.3 we have that the element $v_{\max}$, which maximizes $\mathcal{L}_{\mathcal{C}}^{2,2}(P, v)$, is in $\mathcal{U}_1$. It follows by induction, Equation (4.25) and Theorem 2.3 that

$$\mathcal{L}_{\mathcal{C}}^{2,2}(P) < 1 + 5(1 - p_{\max})p_{\max} + \frac{55}{16}p_{\max}^2.$$

Since the right hand side is monotone increasing in $p_{\max}$ and $p_{\max} \leq 1/2$ we obtain

$$\mathcal{L}_{\mathcal{C}}^{2,2}(P) < 1 + \frac{5}{4} + \frac{55}{16} \cdot \frac{1}{4} = \frac{199}{64} < \frac{55}{16}.$$

In the following, whenever there occurs the case that $p_{\max} \leq 1/2$ we obtain for the same reasons as above that $\mathcal{L}_{\mathcal{C}}^{2,2}(P) < 199/64 < 55/16$.

Case 2: $\mathbf{p_1 < \frac{1}{2}}$

Case 2.1: $\mathbf{t = 1}$
We obtain by the definition of t that $\sum_{u=1}^{4} p_u > 1/2$. If $v_{\max} \in \mathcal{U}_0$ it follows that $\mathcal{L}_{\mathcal{C}}^{2,2}(P) \leq 2$. Further, we get for $v_{\max} \in \mathcal{U}_1$ that $p_{\max} < 1/2$.

Case 2.2: $\mathbf{t \geq 2}$

Case 2.2.1: $\mathbf{v_{\max} \in \mathcal{U}_0}$
We have $p_{\max} = \sum_{u=1}^{t} p_u$. If $t = 2$, we again get that $p_{\max} \leq 1/2$ and if $t = 3$ we get by the same case within Case 2.2.1 of the proof of Theorem 2.3 that

$$\mathcal{L}_{\mathcal{C}_{\max}}^{1,2}(P_{\max}) = 1 + \frac{p_2 + p_3}{p_{\max}} \leq \frac{5}{3}.$$

Further, for the same reasons we obtain

$$\mathcal{L}_{\mathcal{C}_{\max}}^{2,2}(P_{\max}) = 1 + \frac{2(p_2 + p_3)}{p_{\max}} \leq \frac{7}{3}.$$

Applying the above two equations together with Equations (2.7) and (4.25) yields

$$\mathcal{L}_{\mathcal{C}}^{2,2}(P) \leq 1 + 2 \cdot \frac{3}{8} \cdot \frac{5}{8} \cdot \frac{5}{3} + \frac{25}{64} \cdot \frac{7}{3} = \frac{517}{192} < \frac{55}{16}.$$

For $t \geq 4$ we get by Equation (2.7) that

$$p_{\max} < \frac{7}{12}$$

if $p_{\max} \geq 1/2$. This together with the induction hypothesis and Theorem 2.3 yields

$$\mathcal{L}_{\mathcal{C}}^{2,2}(P) < 1 + 5 \cdot \frac{5}{12} \cdot \frac{7}{12} + \frac{49}{144} \cdot \frac{55}{16} = \frac{7799}{2304} < \frac{55}{16}.$$

Case 2.2.2: $\mathbf{v}_{\max} \in \mathcal{U}_1$
We get $p_{\max} = \sum_{u=t+1}^{N} p_u$. Now, Equation (2.8) yields

$$p_{\max} = 1 - \sum_{u=1}^{t} p_u \leq \frac{3}{5}.$$

From this it follows together with the induction hypothesis and Theorem 2.3 that

$$\mathcal{L}_{\mathcal{C}}^{2,2}(P) < 1 + 5 \cdot \frac{2}{5} \cdot \frac{3}{5} + \frac{9}{25} \cdot \frac{55}{16} = \frac{55}{16}.$$

□

We have established a lower and an upper bound for binary 2-identification so that we close this section with

Corollary 4.13 *It holds for all probability distributions P on $\mathcal{U}$ that*

$$4\left(1 - \sum_{u \in \mathcal{U}} p_u^2\right) - \frac{4}{3}\left(1 - \sum_{u \in \mathcal{U}} p_u^3\right) \leq \mathcal{L}^{2,2}(P, P) \leq \mathcal{L}^{2,2}(P) < \frac{55}{16}.$$

5 L-Identification for General Distributions

We now try to generalize the results of the preceding section. We begin with the definition of the q-ary identification entropy of degree L. Again, this function obeys some important desiderata for entropy functions. However, we did not succeed in proving the analogous lower and upper bounds for these entropies. In fact, there exist counterexamples to the natural conjecture that the uniform distribution is an upper bound. In order to show that $H_{\mathrm{ID}}^{L,q}$ is a lower bound for L-identification we only need the bounds for the case where the size of the output space equals the size of the alphabet. We show that we can prove $H_{\mathrm{ID}}^{L,q} \leq \mathcal{L}_{\mathcal{C}}^{L,q}(P,P)$ if we assume that in this case the uniform distribution is indeed an upper bound. Moreover, if we assume that for $N = q$ the uniform distribution is the only distribution for which the upper bound of $H_{\mathrm{ID}}^{L,q}$ is attained, we can show that again if and only if P consists only of q-powers we get that there exists a code $\mathcal{C}$ such that $H_{\mathrm{ID}}^{L,q}(P) = \mathcal{L}_{\mathcal{C}}^{L,q}(P,P)$.

Definition 5.1 *Let $\mathcal{U}$ be a finite set with $|\mathcal{U}| = N$, $L \in \mathbb{N}$, $q \geq 2$ and $P = (p_1, ..., p_N) \in \Gamma_N$. Then the* **q-ary identification entropy of degree** L $H_{ID}^{L,q} : \Gamma_N \to \mathbb{R}$ *is defined by*

$$H_{ID}^{L,q}(P) = -\sum_{l=1}^{L}(-1)^l \binom{L}{l} \frac{q^l}{q^l - 1}\left(1 - \sum_{u\in\mathcal{U}} p_u^{l+1}\right).$$

It is an easy observation that for $L = 1$ the above function equals the identification entropy established in [3]. Also for $L = 2$ it coincides with the identification entropy of second degree from Subsection 4.2.

This function again obeys important desiderata for entropies from [1]. It clearly is symmetric, expansible and decisive. It is also normalized. This follows from

$$H_{\mathrm{ID}}^{L,q}\left(\frac{1}{q}, ..., \frac{1}{q}\right) = -\sum_{l=1}^{L}(-1)^l \binom{L}{l} = 1. \tag{5.1}$$

Another interesting property is that $H_{\mathrm{ID}}^{L,q}$ obeys a grouping behavior which is a generalized version of the grouping behavior of the q-ary identification entropy of the second degree. With the same definitions as in 6. of Theorem 4.9 we obtain

$$H_{\mathrm{ID}}^{L,q}(P) = H_{\mathrm{ID}}^{L,q}(Q) + \sum_{i=1}^{m}\sum_{l=1}^{L}\binom{L}{l}Q_i^{l+1}(1-Q_i)^{L-l}H_{\mathrm{ID}}^{l,q}(P_i). \tag{5.2}$$

To see this we set

$$S_i = \sum_{l=1}^{L}\binom{L}{l}Q_i^{l+1}(1-Q_i)^{L-l}H_{\mathrm{ID}}^{l,q}(P_i),$$

for all $i \in [m]$ and observe that S_i equals

$$\begin{aligned}
& -\sum_{l=1}^{L}\tbinom{L}{l}Q_i^{l+1}(1-Q_i)^{L-l}\sum_{k=1}^{l}(-1)^k\tbinom{l}{k}\tfrac{q^k}{q^k-1}(1-Q_i^{-(k+1)}\sum_{u\in\mathcal{U}_i}p_u^{k+1}) \\
= \; & -\sum_{k=1}^{L}(-1)^k\tfrac{q^k}{q^k-1}(1-Q_i^{-(k+1)}\sum_{u\in\mathcal{U}_i}p_u^{k+1})\sum_{l=k}^{L}\tbinom{L}{l}\tbinom{l}{k}Q_i^{l+1}(1-Q_i)^{L-l} \\
= \; & -\sum_{k=1}^{L}(-1)^k\tbinom{L}{k}\tfrac{q^k}{q^k-1}(Q_i^{k+1}-\sum_{u\in\mathcal{U}_i}p_u^{k+1})\sum_{l=k}^{L}\tbinom{L-k}{l-k}Q_i^{l-k}(1-Q_i)^{L-l} \\
= \; & -\sum_{k=1}^{L}(-1)^k\tbinom{L}{k}\tfrac{q^k}{q^k-1}(Q_i^{k+1}-\sum_{u\in\mathcal{U}_i}p_u^{k+1}).
\end{aligned}$$

Here, the last equality follows from

$$\sum_{l=k}^{L}\binom{L-k}{l-k}Q_i^{l-k}(1-Q_i)^{L-l} = \sum_{l=0}^{L-k}\binom{L-k}{l}Q_i^{l}(1-Q_i)^{L-l-k} = 1.$$

If we now replace k by l, we obtain

$$\sum_{i=1}^{m}S_i = -\sum_{l=1}^{L}(-1)^l\binom{L}{l}\frac{q^l}{q^l-1}(\sum_{i=1}^{m}Q_i^{l+1}-\sum_{u\in\mathcal{U}}p_u^{l+1}).$$

This yields

$$\begin{aligned}
& H_{\mathrm{ID}}^{L,q}(Q) + \sum_{i=1}^{m}S_i \\
= \; & -\sum_{l=1}^{L}(-1)^l\tbinom{L}{l}\tfrac{q^l}{q^l-1}\left(1-\sum_{i=1}^{m}Q_i^{l+1}+\sum_{i=1}^{m}Q_i^{l+1}-\sum_{u\in\mathcal{U}}p_u^{l+1}\right) \\
= \; & H_{\mathrm{ID}}^{L,q}(P).
\end{aligned}$$

The crucial part are the lower and upper bound. It is natural for an entropy function that it is minimized if the probability is 1 for a single object and upper bounded by the uniform distribution. However, we encountered counterexamples such as $L \geq 4$, $q \geq 15$ and $N = 2$ or $L \geq 5$, $q \geq 100$ and $N = 3$. We conjecture that it holds at least for $N \geq q$ and all L and q that

$$H_{\mathrm{ID}}^{L,q}(1, 0, ..., 0) \leq H_{\mathrm{ID}}^{L,q}(P) \leq H_{\mathrm{ID}}^{L,q}\left(\frac{1}{N}, ..., \frac{1}{N}\right). \tag{5.3}$$

This claim, in fact just in the case $N = q$, would suffice to prove that $H_{\mathrm{ID}}^{L,q}$ is a lower bound for L-identification. We did not succeed in proving this claim in general for all L and q and will discuss this problem in greater detail in Section 7. Before we turn to the cases for which we were able to prove the desired bounds, we state

Proposition 5.2 *If Equation (5.3) holds for $N = q$, we get*

$$H_{ID}^{L,q}(P) \leq \mathcal{L}^{L,q}(P, P).$$

Proof:

We will use induction over L and N. As the induction base for L serves the case $L = 1$ for which it has been proven in [3] that identification entropy (of first degree) is a lower bound for (1-)identification. Also the case $L = 2$ has been settled in the preceding Section 4.

The induction base for N is the case $N = q$. By the expansibility property this case settles all necessary induction bases $1, ..., q$. Trivially, if $\mathcal{C} = \mathcal{Q}$, we get that $\mathcal{L}_{\mathcal{C}}^{L,q}(P) = 1$. Since we have assumed that Equation (5.3) holds, Equation (5.1) proves this induction base.

To prove the proposition we partition $\mathcal{U}$ according to some given code $\mathcal{C}$ into $\mathcal{U}_0, ..., \mathcal{U}_{q-1}$, where $\mathcal{U}_i = \{u \in \mathcal{U} \ : \ c_{u,1} = i\}$. Further, let Q be a probability distribution on $\mathcal{Q}$ defined by $Q_i = \sum_{u \in \mathcal{U}} p_u$ and P_i be probability distributions on $\mathcal{U}$ defined by $P_{i,u} = p_u / Q_i$ for all $u \in \mathcal{U}$. With these definitions we obtain

$$\begin{aligned}
\mathcal{L}_{\mathcal{C}}^{L,q}(P, P) &= 1 + \sum_{i \in \mathcal{Q}} \sum_{l=1}^{L} \binom{L}{l} Q_i^{l+1} (1 - Q_i)^{L-l} \mathcal{L}_{\mathcal{C}^{(i)}}^{l,q}(P_i^l, P_i) \\
&\geq H_{\mathrm{ID}}^{L,q}(Q) + \sum_{i=1}^{m} \sum_{l=1}^{L} \binom{L}{l} Q_i^{l+1} (1 - Q_i)^{L-l} H_{\mathrm{ID}}^{l,q}(P_i) \\
&= H_{\mathrm{ID}}^{L,q}(P).
\end{aligned} \tag{5.4}$$

Here the first equality follows from Lemma 4.10, the inequality from the normalization property (5.1), the assumed bounds (5.3) and the induction base. The final equality is a consequence of the grouping behavior (5.2).

□

As stated before there are some cases for which we can prove Equation (5.3). In fact, we prove more, namely

Proposition 5.3 *$H_{ID}^{L,2}(P)$ is strictly concave for $L \leq 20$.*

Proof:

Let

$$f(p) = H_{\mathrm{ID}}^{L,2}(p, 1-p) = -\sum_{l=1}^{L}(-1)^l \binom{L}{l} \frac{2^l}{2^l - 1} \left(1 - p^{l+1} - (1-p)^{l+1}\right).$$

If we now look at all derivatives, we see that for $k = 1$

$$\frac{\delta^k}{\delta^k p} f(p) = \sum_{l=1}^{L}(-1)^l \binom{L}{l} \frac{2^l}{2^l - 1} \frac{(l+1)!}{(l-k+1)!} \left(p^{l-k+1} - (1-p)^{l-k+1}\right)$$

and for all $k \in \{2, ..., L+1\}$

$$\frac{\delta^k}{\delta^k p} f(p) = \sum_{l=k-1}^{L}(-1)^l \binom{L}{l} \frac{2^l}{2^l - 1} \frac{(l+1)!}{(l-k+1)!} \left(p^{l-k+1} + (-1)^k (1-p)^{l-k+1}\right).$$

A first observation is that if k is odd, we get

$$\frac{\delta^k}{\delta^k p} f\left(\frac{1}{2}\right) = 0.$$

If we sort $f(p)$ with respect to the power of p, we get

$$\begin{aligned} f(p) = \quad & ((-1)^L - 1)\tfrac{2^L}{2^L - 1} p^{L+1} \\ & + \left((L+1)\tfrac{2^L}{2^L - 1} - (1 + (-1)^L) L \tfrac{2^{L-1}}{2^{L-1} - 1}\right) p^L + \sum_{l=1}^{L-1} \alpha_l p^l, \end{aligned}$$

for some α_l. This yields that for even L we have a polynomial of degree L with

$$\frac{\delta^L}{\delta^L p} f(p) = \left((L+1)\frac{2^L}{2^L - 1} - 2L\frac{2^{L-1}}{2^{L-1} - 1}\right) L! < 0$$

and for odd L we have a polynomial of degree $L+1$ with

$$\frac{\delta^{L+1}}{\delta^{L+1}p}f(p) = -2\frac{2^L}{2^L-1}(L+1)! < 0.$$

Since for even (resp. odd) L the L-th (resp. $(L+1)$-th) derivative is a strictly negative constant, we know that the $(L-2)$-th (resp. $(L-1)$-th) derivative is a concave function. To show that it is also strictly negative it suffices to show that it is negative for $p = 1/2$ since the $(L-1)$-th (L-th) derivative is zero only at this point. This step can then be iterated and if we can show that all even derivatives are strictly negative at $p = 1/2$, we finally obtain that $H_{\mathrm{ID}}^{L,2}$ is a concave function. For $L = 2, ..., 20$ the values of all even derivatives at $p = 1/2$ have been computed and turn out to be strictly negative.

□

For $L \geq 21$ there occur positive values within the even derivatives so that we cannot prove concavity via this argument. Nevertheless, also for these cases the graphs of the identification entropy functions let us assume that they are still concave. Since the binary identification entropy of degrees up to 20 are concave and symmetric, we obtain

Corollary 5.4 *Let $L \leq 20$ it then holds that*

$$H_{ID}^{L,2}(1,0,...,0) \leq H_{ID}^{L,2}(P) \leq H_{ID}^{L,2}\left(\frac{1}{N},...,\frac{1}{N}\right),$$

with equality on the right hand side if and only if $P = (1/N, ..., 1/N)$.

The cases proved above and especially the strong connection between the grouping behavior (5.2) and Lemma 4.10 provide us with strong believe that the q-ary identification entropy of degree L is indeed a lower bound for the symmetric L-identification running time. But there are two other encouraging facts about the connection between those two concepts. The first is that we get for the uniform distribution the same result like for 2-identification. In fact, we have

$$H_{\mathrm{ID}}^{L,q}\left(\frac{1}{N},...,\frac{1}{N}\right) = -\sum_{l=1}^{L}(-1)^l\binom{L}{l}\frac{q^l}{q^l-1}(1-\frac{1}{N^l})$$

yielding

$$\lim_{N\to\infty} H_{\mathrm{ID}}^{L,q}\left(\frac{1}{N},...,\frac{1}{N}\right) = -\sum_{l=1}^{L}(-1)^l\binom{L}{l}\frac{q^l}{q^l-1}$$

and thus, if $\mathcal{C} \in \mathcal{C}_{q,N}$,

$$\lim_{N\to\infty} H_{\text{ID}}^{L,q}\left(\frac{1}{N},...,\frac{1}{N}\right) = \lim_{N\to\infty} \mathcal{L}_{\mathcal{C}}^{L,q}\left(\frac{1}{N},...,\frac{1}{N}\right).$$

Therefore, a proof of Equation (5.3) would also imply that for the case of the uniform distribution balanced Huffman codes are asymptotically optimal for L-identification.

The second encouraging fact is stated in the following

Proposition 5.5 *Let P be a probability distribution on $\mathcal{U}$ which consists only of q-powers and $\mathcal{C}$ be a code for $(\mathcal{U}, P)$ with $\|c_u\| = -\log_q p_u$ for all $u \in \mathcal{U}$. Then for all L and q it holds that*

$$H_{ID}^{L,q}(P) = \mathcal{L}_{\mathcal{C}}^{L,q}(P,P).$$

Proof:

We first introduce for all $v \in \mathcal{U}$ and $k = 1, ..., \|c_v\|$ the following sets

- $\mathcal{U}_{v,k}^{L} = \{u^L \in \mathcal{U}^L \ : \ \mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v) = k\}$
- $\mathcal{U}_{v,k} = \{u \in \mathcal{U} \ : \ \mathcal{L}_{\mathcal{C}}^{1,q}(u, v) = k\}$
- $\bar{\mathcal{U}}_{v,k} = \mathcal{U}_{v,1} \dot{\cup} ... \dot{\cup} \mathcal{U}_{v,k-1}$

With this notation we obtain

$$\mathcal{L}_{\mathcal{C}}^{L,q}(P,P) = \sum_{v\in\mathcal{U}} p_v \sum_{k=1}^{\|c_v\|} k \sum_{u^L\in\mathcal{U}_{v,k}^L} P_{u^L}^L.$$

We use $S_k = \sum_{u^L\in\mathcal{U}_{v,k}^L} p_{u_1}...p_{u_L}$ and obtain

$$S_k = \sum_{l=1}^{L} \binom{L}{l} \sum_{u_1,...,u_l\in\mathcal{U}_{v,k}} \ \sum_{u_{l+1},...,u_L\in\bar{\mathcal{U}}_{v,k}} p_{u_1}...p_{u_L}.$$

Here, the second equality holds because there has to be at least one output for which identification against v takes exactly k timesteps while all others (or none if $l = L$) have an identification time regarding v of at most $k-1$.

Case 1: $\mathbf{k = 1, ..., \|c_v\| - 1}$
In this case we have that $\mathcal{U}_{v,k} = \bar{T}_{c_v^{k-1}} \backslash \bar{T}_{c_v^k}$ and $\bar{\mathcal{U}}_{v,k} = \bar{T}_{\mathcal{C}} \backslash \bar{T}_{c_v^{k-1}}$. This yields

$$\sum_{u \in \mathcal{U}_{v,k}} p_u = P(T_{c_v^{k-1}}) - P(T_{c_v^k}) = q^{-k+1} - q^{-k} = q^{-k}(q-1)$$

and

$$\sum_{u \in \bar{\mathcal{U}}_{v,k}} p_u = 1 - P(T_{c_v^{k-1}}) = 1 - q^{-k+1}$$

and therewith

$$S_k = \sum_{l=1}^{L} \binom{L}{l} q^{-kl} (q-1)^l (1 - q^{-k+1})^{L-l} = (1 - q^{-k})^L - (1 - q^{-k+1})^L.$$

Case 2: $\mathbf{k = \|c_v\|}$
In this case we have that $\mathcal{U}_{v,\|c_v\|} = \bar{T}_{c_v^{\|c_v\|-1}}$ and $\bar{\mathcal{U}}_{v,\|c_v\|} = \bar{T}_{\mathcal{C}} \backslash \bar{T}_{c_v^{\|c_v\|-1}}$. We obtain

$$\sum_{u \in \mathcal{U}_{v,\|c_v\|}} p_u = P(T_{c_v^{\|c_v\|-1}}) = q^{-\|c_v\|+1}$$

and

$$\sum_{u \in \bar{\mathcal{U}}_{v,\|c_v\|}} p_u = 1 - P(T_{c_v^{\|c_v\|-1}}) = 1 - q^{-\|c_v\|+1}$$

and therewith

$$S_{\|c_v\|} = \sum_{l=1}^{L} \binom{L}{l} q^{-(\|c_v\|-1)l} (1 - q^{-\|c_v\|+1})^{L-l} = 1 - (1 - q^{-\|c_v\|+1})^L.$$

Combining the above two cases yields

$$\begin{aligned}
& \sum_{k=1}^{\|c_v\|} k S_k \\
= & \sum_{k=1}^{\|c_v\|-1} k \left[(1 - q^{-k})^L - (1 - q^{-k+1})^L \right] + \|c_v\| \left[1 - (1 - q^{-\|c_v\|+1})^L \right] \\
= & \sum_{k=1}^{\|c_v\|-1} k (1 - q^{-k})^L + \|c_v\| - \sum_{k=1}^{\|c_v\|} k (1 - q^{-k+1})^L.
\end{aligned}$$

We set $A = \sum_{k=1}^{\|c_v\|-1} k(1-q^{-k})^L + \|c_v\|$ and $B = \sum_{k=1}^{\|c_v\|} k(1-q^{-k+1})^L$. We then get for A

$$\begin{aligned}
A = & \sum_{k=1}^{\|c_v\|-1} k \sum_{l=0}^{L} \binom{L}{l} (-1)^{L-l} q^{-(L-l)k} + \|c_v\| \\
= & \sum_{l=0}^{L} \binom{L}{l} (-1)^{L-l} \sum_{k=1}^{\|c_v\|-1} k q^{-(L-l)k} + \|c_v\| \\
= & \sum_{l=0}^{L-1} \binom{L}{l} (-1)^{L-l} \sum_{k=1}^{\|c_v\|-1} k q^{-(L-l)k} + \sum_{k=1}^{\|c_v\|} k \\
= & \sum_{l=0}^{L-1} \binom{L}{l} (-1)^{L-l} \sum_{k=1}^{\|c_v\|} k q^{-(L-l)k} \\
& - \sum_{l=0}^{L-1} \binom{L}{l} (-1)^{L-l} \|c_v\| q^{-(L-l)\|c_v\|} + \sum_{k=1}^{\|c_v\|} k
\end{aligned}$$

and for B respectively

$$\begin{aligned}
B = & \sum_{k=1}^{\|c_v\|} k \sum_{l=0}^{L} \binom{L}{l} (-1)^{L-l} q^{-(L-l)(k-1)} \\
= & \sum_{l=0}^{L} \binom{L}{l} (-1)^{L-l} q^{L-l} \sum_{k=1}^{\|c_v\|} k q^{-(L-l)k} \\
= & \sum_{l=0}^{L-1} \binom{L}{l} (-1)^{L-l} q^{L-l} \sum_{k=1}^{\|c_v\|} k q^{-(L-l)k} + \sum_{k=1}^{\|c_v\|} k.
\end{aligned}$$

Subtracting B from A yields

$$\begin{aligned}
\sum_{k=1}^{\|c_v\|} k S_k = & \sum_{l=0}^{L-1} \binom{L}{l} (-1)^{L-l} (1-q^{L-l}) \sum_{k=1}^{\|c_v\|} k q^{-(L-l)k} \\
& - \sum_{l=0}^{L-1} \binom{L}{l} (-1)^{L-l} \|c_v\| q^{-(L-l)\|c_v\|}.
\end{aligned}$$

Since

$$\sum_{k=1}^{\|c_v\|} kq^{-(L-l)k} = \frac{q^{L-l}}{(q^{L-l}-1)^2}(1-q^{-(L-l)\|c_v\|}) - \frac{\|c_v\|q^{-(L-l)\|c_v\|}}{q^{L-l}-1}$$

and by assumption of the theorem $q^{-\|c_v\|} = p_v$ for all $v \in \mathcal{U}$, we finally obtain

$$\mathcal{L}_{\mathcal{C}}^{L,q}(P,P) = \sum_{v\in\mathcal{U}} p_v \sum_{k=1}^{\|c_v\|} kS_k = -\sum_{l=1}^{L} \binom{L}{l}(-1)^l \frac{q^l}{q^l-1}(1-\sum_{v\in\mathcal{U}} p_v^{l+1}) = H_{\mathrm{ID}}^{L,q}(P).$$

□

According to the previous results for (1-) and 2-identifications it seems natural that the equality of Proposition 5.5 is only assumed for the mentioned cases and that we have a strict inequality between the q-ary identification entropy of degree L and the symmetric L-identification running time if P does not consists only of q-powers. The following proposition formalizes this if we assume that for $N = q$ the uniform distribution maximizes $H_{\mathrm{ID}}^{L,q}$ and that

$$H_{\mathrm{ID}}^{L,q}(P') < H_{\mathrm{ID}}^{L,q}\left(\frac{1}{q},...,\frac{1}{q}\right)$$

for all other distributions $P' \neq (1/q,...,1/q)$.

Proposition 5.6 *Let P be a probability distribution on $\mathcal{U}$ for which it holds that*

$$H_{ID}^{L,q}(P) = \mathcal{L}_{\mathcal{C}}^{L,q}(P,P).$$

We further assume that $H_{ID}^{L,q}(P') < H_{ID}^{L,q}((1/q,...,1/q))$ for all $P' \neq (1/q,...,1/q)$. It then follows that P consists only of q-powers and $\mathcal{C}$ is a code for $(\mathcal{U},P)$ with $\|c_u\| = -\log_q p_u$ for all $u \in \mathcal{U}$.

Proof:

As induction base serves the case $L = 1$, which has been proven in Theorem 1 in [5]. For the induction steps it now follows from the assumptions of that the inequality in equation 5.4 becomes an equality so that we have

$$\begin{aligned} & 1 + \sum_{i\in\mathcal{Q}} \sum_{l=1}^{L} \tbinom{L}{l} Q_i^{l+1}(1-Q_i)^{L-l} \mathcal{L}_{\mathcal{C}^{(i)}}^{l,q}(P_i^l, P_i) \\ = \quad & H_{\mathrm{ID}}^{L,q}(Q) + \sum_{i=1}^{m} \sum_{l=1}^{L} \tbinom{L}{l} Q_i^{l+1}(1-Q_i)^{L-l} H_{\mathrm{ID}}^{l,q}(P_i). \end{aligned} \tag{5.5}$$

For the definitions of Q_i, P_i and $\mathcal{C}^{(i)}$ see again Lemma 4.10. From this equation follows

i) $H_{\mathrm{ID}}^{L,q}(Q) = 1$

ii) $H_{\mathrm{ID}}^{l,q}(P_i) = \mathcal{L}_{\mathcal{C}^{(i)}}^{l,q}(P_i^l, P_i)$ for $l \in [L]$

On the one hand it follows from the assumptions and i) that $Q = (1/q, ..., 1/q)$ and on the other hand it follows from the induction hypothesis and ii) that P_i consists only of q-powers and that $\|c_u^{(i)}\| = -\log_q p_{i,u}$. Since $p_u = Q_i p_{i,u} = p_{i,u}/q$ for all $u \in \mathcal{U}_i$, we obtain that also P consists only of q-powers and finally $\|c_u\| = -\log_q p_{i,u} + 1 = -\log_q \frac{p_{i,u}}{q} = -\log_q p_u$ for all $u \in \mathcal{U}_i$.

□

6 L-Identification for Sets

Like before the discrete source $(\mathcal{U}, P)$ together with a source code $\mathcal{C}$ forms the basis for our analysis of L-identification for sets. Unlike Subsection, however, 1.2 we do not consider as the output space the discrete memoryless source $(\mathcal{U}^L, P^L)$ but the discrete source $(\tilde{\mathcal{U}}, \tilde{P})$, where $\tilde{\mathcal{U}} = \binom{\mathcal{U}}{L}$. We write $\tilde{P}_S$ for $\tilde{P}(\{S\})$. The task of L-identification for sets is in principle the same as before. It has to be able to distinguish for all users $v \in \mathcal{U}$ and all outputs $S \in \tilde{\mathcal{U}}$ whether there exists an element u in S with $u = v$ or not.

In this section we will analyze the asymptotic behavior of the symmetric running time of L-identification for sets for the case when $\tilde{P}$ is the uniform distribution on $\tilde{\mathcal{U}}$ and also the users are chosen uniformly. We will see that it asymptotically equals the symmetric running time of L-identification (for vectors) and thus $K_{L,q}$, which was examined in Subsection 3.2.

It is clear that L-identification for sets can be seen as a special case of our preliminary L-identification (for vectors) as we exclude all vectors with two or more identical components. This fact changes the running time of L-identification in the following way. Again, we compare q-bit by q-bit the codewords of the elements of S to the corresponding q-bit of c_v and after every step we cancel out all elements which do not coincide. Suppose after some step k during the identification process we are left over with the same amount of possible candidates as there are codewords in $\bar{\mathcal{N}}(T_{c_v^k})$. Since we are considering sets and not vectors, we know that each of the elements which belong to the codewords in $\bar{\mathcal{N}}(T_{c_v^k})$ are elements of S and so does v itself. At such a point we terminate the identification process and answer: "Yes, v is in S!". Figure 6.1 shows an example of such an event for $N = 17$ and $L = 9$. In this example v equals u_1. This is indicated by the thick path from the root to u_1. After the first q-ary comparison u_5 and u_7 are deleted from the set of possible candidates but there are more than seven codewords which begin with $\mathbf{0}$ so that v still might be not contained in S. After the second comparison u_2 and u_9 are canceled and we still have more codewords in $\bar{\mathcal{N}}(T_{\mathbf{00}})$ than possible candidates. After the third step, however, u_6 is not longer a candidate. This leaves us with four possible candidates. Since $|\bar{\mathcal{N}}(T_{\mathbf{000}})| = 4$, we know that v has to be an element of S and terminate the L-identification process.

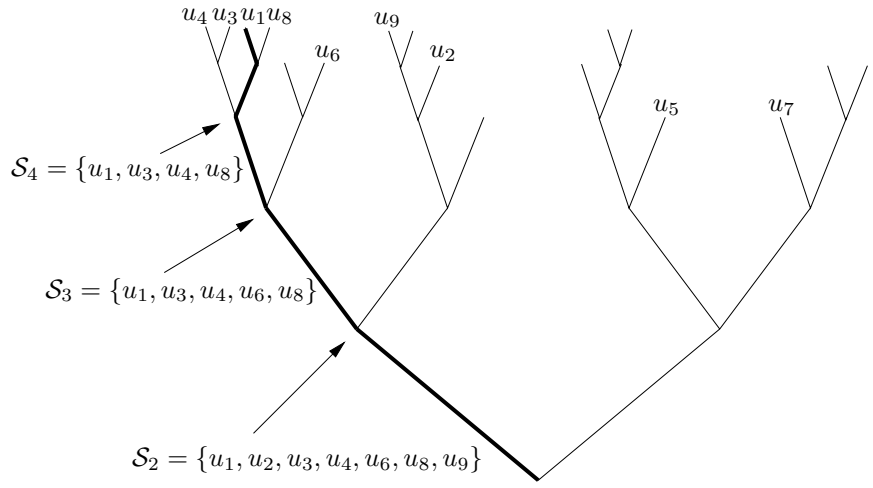

Figure 6.1: An example when the 9-identification process terminates because $|\mathcal{S}_4| = |\{u \in \mathcal{U} \ : \ c_u^3 = c_v^3\}| = 4$. For the definition of $\mathcal{S}_i$ see Table 8.2 in the appendix.

The L-identification algorithm `LID` now becomes the L-identification algorithm for sets. It is called `LIDforSets` and stated in Table 8.2 in the appendix. Now let $S = \{u_1, u_2, ..., u_L\} \in \tilde{\mathcal{U}}$ we then define the L-identification time for S, an user v and a q-ary code $\mathcal{C}$ by

$$\tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}(S, v) = \text{LIDforSets}_2(c_{u_1}, ..., c_{u_L}, c_v), \tag{6.1}$$

where $\text{LIDforSets}_2(c_{u_1}, ..., c_{u_L}, c_v)$ is the second component of the return pair of the algorithm `LIDforSets`.

In the same way as in Subsection 1.2 we now define the *average running time* for a given user $v \in \mathcal{U}$ by

$$\tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}(\tilde{P}, v) = \sum_{S \in \tilde{\mathcal{U}}} \tilde{P}_S \tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}(S, v), \tag{6.2}$$

the *worst-case running-time* by

$$\tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}(\tilde{P}) = \max_{v \in \mathcal{U}} \tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}(\tilde{P}, v) \tag{6.3}$$

and if we have a probability distribution Q on $\mathcal{U}$, we define the *expected running time* by

$$\tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}(\tilde{P}, Q) = \sum_{v \in \mathcal{U}} Q(\{v\}) \tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}(\tilde{P}, v).^1 \tag{6.4}$$

[1]Remember that all those functions implicitly depend also on $N = |\mathcal{U}|$ via $\mathcal{C}$, $\tilde{P}$ and Q.

In both scenarios we are again interested in the optimal running time. That is

$$\tilde{\mathcal{L}}^{L,q}(\tilde{P}) = \min_{\mathcal{C}} \tilde{\mathcal{L}}^{L,q}_{\mathcal{C}}(\tilde{P}) \tag{6.5}$$

and

$$\tilde{\mathcal{L}}^{L,q}(\tilde{P}, Q) = \min_{\mathcal{C}} \tilde{\mathcal{L}}^{L,q}_{\mathcal{C}}(\tilde{P}, Q). \tag{6.6}$$

We will now take a look at the asymptotic behavior of $\tilde{\mathcal{L}}^{L,q}_{\mathcal{C}}(\tilde{P}, Q)$ for the case when both $\tilde{P}$ and Q are uniform distributions on $\tilde{\mathcal{U}}$, resp. $\mathcal{U}$, and that $\mathcal{C} \in \mathcal{C}_{q,N}$ is a balanced Huffman code. In this case we call $\tilde{\mathcal{L}}^{L,q}_{\mathcal{C}}(\tilde{P}, Q)$ as before the *symmetric running time* for L-identification for sets. In order to simplify notation we shall write $\bar{P} = \left(\binom{N}{L}^{-1}, ..., \binom{N}{L}^{-1}\right)$. Equation 6.4 then becomes

$$\tilde{\mathcal{L}}^{L,q}_{\mathcal{C}}\left(\bar{P}, (\frac{1}{N}, ..., \frac{1}{N})\right) = \frac{1}{N\binom{N}{L}} \sum_{S \in \tilde{\mathcal{U}}} \sum_{v \in \mathcal{U}} \tilde{\mathcal{L}}^{L,q}_{\mathcal{C}}(S, v). \tag{6.7}$$

It turns out that

$$\lim_{N\to\infty} \tilde{\mathcal{L}}^{L,q}_{\mathcal{C}}\left(\bar{P}, (\frac{1}{N}, ..., \frac{1}{N})\right) = \lim_{N\to\infty} \mathcal{L}^{L,q}_{\mathcal{C}}\left((\frac{1}{N}, ..., \frac{1}{N}), (\frac{1}{N}, ..., \frac{1}{N})\right) \tag{6.8}$$

and thus equals the same rational number $K_{L,q}$ which has been examined in Subsection 3.2. This may be somewhat surprising at first glance since the output spaces $\mathcal{U}^L$ and $\tilde{\mathcal{U}}$ as well as the underlying algorithms differ from each other. Yet, it becomes clear if we take into account that these differences "disappear" if N goes to infinity. By this we mean that the cardinality of the family of sets, which cause the algorithm `LIDforSets` to terminate with a positive answer before it reaches the last step, is so small that its probability goes to zero as N tends to infinity. The same is true for the set of all vectors which have more than one identical component. We will now formalize the above explanations in order to prove Equation (6.8).

Let $f : \mathcal{U}^L \to \bigcup_{l=1}^{L} \binom{\mathcal{U}}{l}$ be defined by $f(u^L) = \bigcup_{i=1}^{L} \{u_i\}$. Further, let $\mathcal{U}' \subset \mathcal{U}$ be the set of all vectors whose components of are pairwise distinct. It follows that the restriction $f|_{\mathcal{U}'}$ is a surjective mapping from $\mathcal{U}'$ onto $\tilde{\mathcal{U}}$ and that $|f^{-1}(S)| = L!$ for all $S \in \tilde{\mathcal{U}}$. This yields $|\mathcal{U}'| = L!\binom{N}{L}$ and

$$\begin{aligned} & \mathcal{L}^{L,q}_{\mathcal{C}}\left((\tfrac{1}{N}, ..., \tfrac{1}{N}), (\tfrac{1}{N}, ..., \tfrac{1}{N})\right) \\ = \; & \tfrac{1}{N^{L+1}} \sum_{v \in \mathcal{U}} \left[\sum_{u^L \in \mathcal{U}'} \mathcal{L}^{L,q}_{\mathcal{C}}(u^L, v) + \sum_{u^L \in \mathcal{U} \setminus \mathcal{U}'} \mathcal{L}^{L,q}_{\mathcal{C}}(u^L, v)\right]. \end{aligned} \tag{6.9}$$

Since $\sum_{u^L \in \mathcal{U} \setminus \mathcal{U}'} \mathcal{L}^{L,q}_{\mathcal{C}}(u^L, v) \leq (1 + \log_q N) L! \binom{N}{L}$, it follows that the second summand multiplied by $1/N^L$ tends to zero for $N \to \infty$.

We now turn to $\tilde{\mathcal{U}}$ and assume that $N = q^n$ such that $\mathcal{C} = \mathcal{C}_{q^n}$.[2] We define $\tilde{\mathcal{U}}' \subset \tilde{\mathcal{U}}$ to be the family of sets S for which there exists at least one leaf in each subtree with root in level $n-1$ which is not contained in S. We use $T = T_{\mathcal{C}_{q^n}}$ and obtain

$$\tilde{\mathcal{U}}' = \{S \in \tilde{\mathcal{U}} \ : \ \bar{\mathcal{N}}(T_x) \backslash S \neq \emptyset \ \forall x \in \mathcal{Q}^{n-1}\}.$$

It follows that from the nature of the algorithms `LID` and `LIDforSets` that for all $v \in \mathcal{U}$, $S \in \tilde{\mathcal{U}}'$ and $u^L \in f^{-1}(S)$ we have that

$$\tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}(S, v) = \mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v). \tag{6.10}$$

It is clear that if $L < q$, we get that $\tilde{\mathcal{U}}' = \tilde{\mathcal{U}}$ and if $L \geq q$, we obtain that

$$\tilde{\mathcal{U}} \backslash \tilde{\mathcal{U}}' = \bigcup_{x \in \mathcal{Q}^{n-1}} \left(\bar{\mathcal{N}}(T_x) \cup \binom{\tilde{\mathcal{U}} \backslash \bar{\mathcal{N}}(T_x)}{L-q} \right).$$

From this follows that

$$\begin{aligned} |\tilde{\mathcal{U}} \backslash \tilde{\mathcal{U}}'| \ &\leq \textstyle\sum_{x \in \mathcal{Q}^{n-1}} \left| \left(\bar{\mathcal{N}}(T_x) \cup \binom{\tilde{\mathcal{U}} \backslash \bar{\mathcal{N}}(T_x)}{L-q} \right) \right| \\ &= q^{n-1} \left(q + \tbinom{N-q}{L-q} \right) = N + \tfrac{N}{q} \tbinom{N-q}{L-q}. \end{aligned}$$

This yields

$$\begin{aligned} & \tfrac{1}{N\binom{N}{L}} \textstyle\sum_{v \in \mathcal{U}} \sum_{S \in \tilde{\mathcal{U}} \backslash \tilde{\mathcal{U}}'} \tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}(S, v) \\ \leq \ & \tfrac{1}{\binom{N}{L}} \log_q N |\tilde{\mathcal{U}} \backslash \tilde{\mathcal{U}}'| \\ \leq \ & \tfrac{1}{\binom{N}{L}} \log_q N \left(N + \tfrac{N}{q} \tbinom{N-q}{L-q} \right). \end{aligned}$$

The right hand side of the third line tends to zero as N goes to infinity. We return to L-identification for vectors and similar to the definition of $\tilde{\mathcal{U}}'$ we define

$$\mathcal{U}'' = \{u^L \in \mathcal{U}' \ : \ \forall x \in \mathcal{Q}^{n-1} \ \exists \ w \in \bar{\mathcal{N}}(T_x) \text{ and } l \in [L] \text{ s.th. } w \neq u_l\}$$

and for similar reasons as above we obtain that for $N \to \infty$

$$\frac{1}{N^{L+1}} \sum_{v \in \mathcal{U}} \sum_{u^L \in \mathcal{U}' \backslash \mathcal{U}''} \mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v) \to 0.$$

[2] The analysis for $N \neq q^n$, which we omit, involves the same calculations but includes some more case distinctions.

Finally, we can partition $\mathcal{U}'' = \bigcup_{S \in \tilde{\mathcal{U}}'} f^{-1}(S)$ and get

$$\begin{aligned} & \tfrac{1}{N^{L+1}} \sum_{v \in \mathcal{U}} \sum_{u^L \in \mathcal{U}''} \mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v) \\ = \; & \tfrac{1}{N^{L+1}} \sum_{v \in \mathcal{U}} \sum_{S \in \tilde{\mathcal{U}}'} \sum_{u^L \in f^{-1}(S)} \mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v) \\ = \; & \tfrac{L!}{N^{L+1}} \sum_{v \in \mathcal{U}} \sum_{S \in \tilde{\mathcal{U}}'} \tilde{\mathcal{L}}_{\mathcal{C}}^{L,q}(S, v), \end{aligned}$$

where the last equality follows from Equation (6.10). Since $L!/N^L$ asymptotically equals $1/\binom{N}{L}$, we finally proved

Theorem 6.1 *Let $L, n \in \mathbb{N}$, $q \in \mathbb{N}_{\geq 2}$, $q^{n-1} < N \leq q^n$, $\mathcal{C} \in \mathcal{C}_{q,N}$ and $\bar{P}$ be the uniform distribution on $\tilde{\mathcal{U}}$. Then it holds that*

$$\lim_{N \to \infty} \tilde{\mathcal{L}}_{\mathcal{C}}^{L,q} \left(\bar{P}, (\frac{1}{N}, ..., \frac{1}{N}) \right) = \lim_{N \to \infty} \mathcal{L}_{\mathcal{C}}^{L,q} \left((\frac{1}{N}, ..., \frac{1}{N}), (\frac{1}{N}, ..., \frac{1}{N}) \right) = K_{L,q},$$

where $K_{L,q} \in \mathbb{R}$ is defined in Theorem 3.4.

7 Open Problems

In this final section we will give an overview of some open problems which arose during the study of L-identification. We begin with three types of problems concerning L-identification (for vectors). The first is to settle the induction base in the proof of Proposition 5.2. It is the only fragment left in order to completely prove that q-ary identification entropy $H_{\mathrm{ID}}^{L,q}$ of degree L is a lower bound for L-identification.

The second problem is a generalization of Lemmas 2.1 and 4.3 where we proved that concerning block codes the uniform distribution is optimal for (1-) and 2-identification. At least for $L \geq 4$ this is not longer true in general as there exist simple counterexamples. However, we claim that if the size of the block is sufficiently large, again uniform distribution becomes optimal.

The second subsection covers L-identification for sets. We have seen in Subsection 6 that for the uniform distribution L-identification for sets behaves in the same way as L-identification (for vectors) if the cardinality of the output space tends to infinity. Unfortunately we have not made any major discoveries if we turn to general distributions.

7.1 Some Open Problems for L-Identification

Induction Base for the proof of Proposition 5.2

The most important problem is to settle for all L and q the induction base $N = q$ of the proof of Proposition 5.2. With the solution of this problem we would obtain that the q-ary identification entropy $H_{\mathrm{ID}}^{L,q}$ of degree L is a lower bound for L-identification. In the following we establish a chain of problems which are partly subproblems. Figure 7.1 visualizes this chain.

Problem 1:
Show that it holds for all L, q and probability distributions P on $[q]$ that

$$H_{\mathrm{ID}}^{L,q}(P) \leq 1. \tag{7.1}$$

Since $H_{\mathrm{ID}}^{L,q}$ is normalized (see Equation (5.1)), the above problem is equivalent to

Problem 1*:
Show that it holds for all L, q and probability distributions P on $[q]$ that

$$H_{\mathrm{ID}}^{L,q}(P) \leq H_{\mathrm{ID}}^{L,q}\left(\frac{1}{q},...,\frac{1}{q}\right). \tag{7.2}$$

We have claimed in Section 5 that Equation (5.3) holds which solves problem 1^* in the more general form where $N \geq q$. This yields

Problem 1.1:
Show that it holds for all L, q and probability distributions P on $[N]$, where $N \geq q$, that

$$H_{\mathrm{ID}}^{L,q}(1,0,...,0) \leq H_{\mathrm{ID}}^{L,q}(P) \leq H_{\mathrm{ID}}^{L,q}\left(\frac{1}{N},...,\frac{1}{N}\right).$$

We provide three approaches which possibly are suitable for solving problem 1.1. The first is somewhat in the spirit of Lemmas 2.1 and 4.3 where we step by step adjust an arbitrary probability distribution so that it becomes the uniform distribution without increasing the symmetric L-identification running time. For this let $P \neq (1/N,...,1/N)$ be a probability distribution on $[N]$. Remember that we assumed $N \geq q$. Clearly, there exists an element, say 1, for which $p_1 > 1/N$ and an element, say 2, for which $p_2 < 1/N$. We now construct a new probability distribution $\bar{P}$ by setting $\bar{p}_1 = \bar{p}_2 = (p_1 + p_2)/2$ and $\bar{p}_i = p_i$, for all $i \in \{3,...,N\}$. If we can show that $H_{\mathrm{ID}}^{L,q}(\bar{P}) - H_{\mathrm{ID}}^{L,q}(P) \geq 0$, we would have solved problem 1.1 since we can come arbitrarily close to $(1/N,...,1/N)$ by applying the above construction iteratively and sufficiently many times. Thus we state

Problem 1.1.1:
Show that it holds for all L, q and probability distributions P on $[N]$, where $N \geq q$, that

$$H_{\mathrm{ID}}^{L,q}(\bar{P}) - H_{\mathrm{ID}}^{L,q}(P) \geq 0,$$

where $\bar{P}$ is defined by $\bar{p}_1 = \bar{p}_2 = (p_1 + p_2)/2$ and $\bar{p}_i = p_i$ for all $i \in \{3,...,N\}$.

We begin the calculation of this difference and obtain

$$\begin{aligned} & H_{\mathrm{ID}}^{L,q}(\bar{P}) - H_{\mathrm{ID}}^{L,q}(P) \\ = & \sum_{l=1}^{L}(-1)^l\tbinom{L}{l}\tfrac{q^l}{q^l-1}\left(\tfrac{1}{2^l}(p_1+p_2)^{l+1} - p_1^{l+1} - p_2^{l+1}\right) \\ = & \sum_{l=1}^{L}(-1)^l\tbinom{L}{l}\sum_{t\geq 0} q^{-tl}\left(\tfrac{1}{2^l}(p_1+p_2)^{l+1} - p_1^{l+1} - p_2^{l+1}\right) \\ = & \sum_{i=1}^{2} p_i \sum_{t\geq 0}\left[\left(1-\tfrac{p_1+p_2}{2q^t}\right)^L - (1-\tfrac{p_i}{q^t})^L\right]. \end{aligned}$$

Note that while the first summand is positive the second one is negative. Yet the positive summand is weighted by p_1 which is greater than p_2 by which the negative summand is weighted. We therefore feel that the following problem may be a good candidate for solving the main problem 1. One has to keep in mind that $N \geq q$ is crucial so this fact has to come in play.

Problem 1.1.1.1:
Show that if $N \geq q$, $p_1 + p_2 \leq 1$ and $p_1 > 1/N > p_2$, we get that

$$\sum_{i=1}^{2} p_i \sum_{t\geq 0}\left[\left(1-\frac{p_1+p_2}{2q^t}\right)^L - (1-\frac{p_i}{q^t})^L\right] \geq 0.$$

We also could try to prove problem 1.1 via the direct way. For this consider an probability distribution P on $[N]$ (still $N \geq q$) and assume w.l.o.g. that

$$p_1 \geq p_2 \geq ... \geq p_{n_1} > \frac{1}{N} > p_{n_1+1} \geq ... \geq p_{n_2} \tag{7.3}$$

and $p_{n_2+1} = ... = p_N = 1/N$. With the same calculations as above we obtain

$$H_{\mathrm{ID}}^{L,q}\left(\frac{1}{N},...,\frac{1}{N}\right) - H_{\mathrm{ID}}^{L,q}(P) = \sum_{i=1}^{n_2} p_i \sum_{t\geq 0}\left[\left(1-\frac{1}{Nq^t}\right)^L - (1-\frac{p_i}{q^t})^L\right].$$

Again the first n_1 summands are positive and weighted by the greater weights $p_1, ..., p_{n_1}$. We obtain

Problem 1.1.2:
Show that if $N \geq q$ and if $(p_1, ..., p_N)$ obeys Equation (7.3), we get that

$$\sum_{i=1}^{n_2} p_i \sum_{t\geq 0}\left[\left(1-\frac{1}{Nq^t}\right)^L - (1-\frac{p_i}{q^t})^L\right] \geq 0.$$

Another approach would be to follow the proof of the bounds for the q-ary identification entropy of second degree (see Theorem 4.9). In this proof we analyze the first derivative of the entropy function and showed that there exists only one extremal point namely a maximum at $(1/N, ..., 1/N)$. As we have mentioned in the definition section we only have to consider $N-1$ partial derivatives and obtain for $v \in [N-1]$

$$\frac{\delta}{\delta p_v} H_{\mathrm{ID}}^{L,q} = \sum_{l=1}^{L} (-1)^l \binom{L}{l} \frac{q^l}{q^l - 1} (l+1) \left(p_v^l - (1 - \sum_{u=1}^{N-1} p_u)^l \right).$$

This obviously is zero if $p_1 = ... = p_{N-1} = 1/N$. We are left with

Problem 1.1.3:
Show that $p_1 = ... = p_{N-1} = 1/N$ is the only point in Δ_{N-1} which is for all $v \in [N-1]$ the root of

$$\sum_{l=1}^{L} (-1)^l \binom{L}{l} \frac{q^l}{q^l - 1} (l+1) \left(p_v^l - (1 - \sum_{u=1}^{N-1} p_u)^l \right).$$

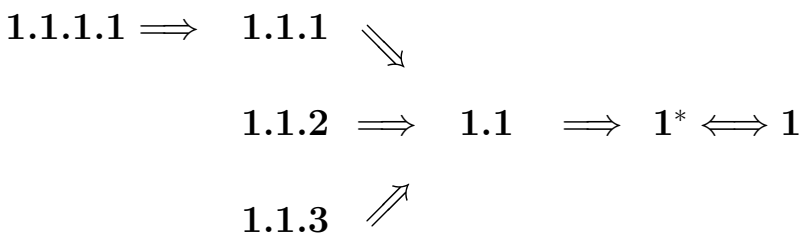

Figure 7.1: The logical chain of the problems leading to a proof of Proposition 5.2.

L-Identification for Block Codes

In Subsections 2.1 and 4.4 we proved that concerning block codes the uniform distribution is optimal for the symmetric running time of (1-) and 2-identification. This, however, is not longer true at least for $L \geq 4$. This we can show by an easy example. Therefore consider $q = 2$, $N = 4$, $L = 4$ and $\mathcal{C} = \mathcal{C}_{2^2}$. It follows with the notation of Subsection 3.2 that

$$\begin{aligned}
& \mathcal{L}_{\mathcal{C}}^{4,2}\left(\left(\tfrac{1}{4},\tfrac{1}{4},\tfrac{1}{4},\tfrac{1}{4}\right),\left(\tfrac{1}{4},\tfrac{1}{4},\tfrac{1}{4},\tfrac{1}{4}\right)\right) \\
= \;& \tfrac{1}{4^4}\left(|\mathcal{R}_{\mathcal{C}}^{4,2}(1,1)| + 2|\mathcal{R}_{\mathcal{C}}^{4,2}(2,1)|\right) \\
= \;& \tfrac{1}{4^4}\left(2^4 + 22^4(2^4-1)\right) \\
= \;& \tfrac{31}{16}.
\end{aligned}$$

We now take the probability distribution $P = (1/8, 1/8, 3/8, 3/8)$. The assignment of the individual probabilities to the codewords (resp. the corresponding outputs) is depicted in Figure 7.2. We obtain

$$\begin{aligned}
& \mathcal{L}_{\mathcal{C}}^{4,2}\left(\left(\tfrac{1}{8},\tfrac{1}{8},\tfrac{3}{8},\tfrac{3}{8}\right),\left(\tfrac{1}{8},\tfrac{1}{8},\tfrac{3}{8},\tfrac{3}{8}\right)\right) \\
= \;& \sum_{i=1}^{2} \sum_{v=2(i-1)+1}^{2i} p_v \sum_{l=0}^{4} \binom{4}{l} \sum_{u_1,..,u_l=2(i-1)+1}^{2i} \sum_{u_{l+1},...,u_4 \in [4]\setminus\{2(i-1)+1,2i\}} P_{u^4}\mathcal{L}(u^4,v) \\
= \;& \tfrac{1}{4}\left(\tfrac{3^4}{2^8} + \tfrac{1}{2^7}\sum_{l=1}^{4}\binom{4}{l}3^{4-l}\right) + \tfrac{3}{4}\left(\tfrac{1}{2^8} + \tfrac{1}{2^7}\sum_{l=1}^{4}\binom{4}{l}3^{l}\right) \\
= \;& \tfrac{491}{256} < \tfrac{496}{256} = \tfrac{31}{16} = \mathcal{L}_{\mathcal{C}}^{4,2}(\tfrac{1}{4},...,\tfrac{1}{4}).
\end{aligned}$$

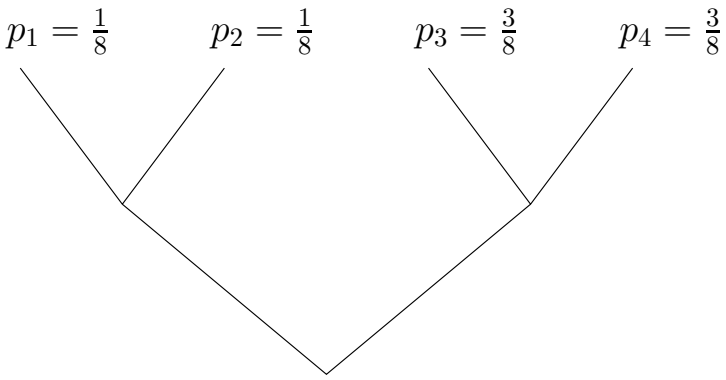

Figure 7.2: An example for 4-identification on block codes which has a faster symmetric running time than the uniform distribution.

This inconsistency disappears for 4-identification already for the next level, where $N = 8$. In general we claim for all L that if the block code is large enough, the uniform distribution becomes optimal again. This is the content of

Problem 2:
Show that for all L exists $n_L \in \mathbb{N}$ such that it holds for all $n \geq n_L$ and all probability distributions P on $[q^n]$ that

$$\mathcal{L}^{L,q}_{\mathcal{C}_{q^n}}(P,P) \geq \mathcal{L}^{L,q}_{\mathcal{C}_{q^n}}\left((\frac{1}{q^n},...,\frac{1}{q^n}),(\frac{1}{q^n},...,\frac{1}{q^n})\right).$$

Of course, we cannot solve this problem by applying generalized versions of Lemmas 2.1 and 4.3. Since these lemmas are applied to small subtrees in the beginning, we would get that during the first modifications of some given probability the symmetric running time would increase if we level out the corresponding probabilities. But we think that these small increases are absorbed by later steps where we level out bigger and bigger subtrees. A big help in order to solve problem 2 would be if we could establish an exact expression for the differences like we have done in Lemma 2.1. With this we would hopefully be able to solve problem 2. However, already for $L = 2$ we do not have such an expression. Like before in the corresponding lemmas for (1-) and 2-identification let $n \in \mathbb{N}$, $q \in \mathbb{N}_{\geq 2}$, $k \in \{0,...,n-1\}$ and $t \in \{0,...,q^{n-k-1}-1\}$. Further, let $P = (p_1,...,p_{q^n})$ and $\tilde{P} = (\tilde{p}_1,...,\tilde{p}_{q^n})$ be probability distributions on $[q^n]$ with

$$P = (p_1,...,p_{tq^{k+1}},\underbrace{r_1,...,r_1}_{q^k},\underbrace{r_2,...,r_2}_{q^k},...,\underbrace{r_q,...,r_q}_{q^k},p_{(t+1)q^{k+1}+1},...,p_{q^n})$$

and

$$\tilde{P} = (p_1,...,p_{tq^{k+1}},\underbrace{\frac{1}{q}\sum_{i=1}^{q} r_i,...,\frac{1}{q}\sum_{i=1}^{q} r_i}_{q^{k+1}},p_{(t+1)q^{k+1}+1},...,p_{q^n}).$$

Problem 2.1:
Establish for $L \geq 2$ an exact expression for the difference

$$\mathcal{L}^{L,q}_{\mathcal{C}_{q^n}}(P,P) - \mathcal{L}^{L,q}_{\mathcal{C}_{q^n}}(\tilde{P},\tilde{P}).$$

7.2 L-Identification for Sets for General Distributions

The basic problem if we turn to general distributions is that the connection between a probability distribution P on $\mathcal{U}$ and a distribution $\tilde{P}$ on $\tilde{\mathcal{U}} = \binom{\mathcal{U}}{L}$ is not as straight forward as it is if we consider the discrete memoryless source $(\mathcal{U}^L, P^L)$, where the probability of a vector is the product of the probabilities of its components. In order to establish such a connection we provide

Definition 7.1 *Let P be a probability distribution on $\mathcal{U}$. Then we define its correlated distribution $P^{(L)}$ on $\tilde{\mathcal{U}}$ by setting*

$$P_S^{(L)} = \sum_{\pi \in \Pi_L} \prod_{l=1}^{L} \frac{p_{s_{\pi(l)}}}{1 - \sum_{m=1}^{l-1} p_{s_{\pi(m)}}}$$

for all $S = \{s_1, ..., s_L\} \in \tilde{\mathcal{U}}$ and where Π_L is the set of all permutations on $[L]$.

This probability equals the probability of a set S which is filled step by step with elements from $\mathcal{U}$ according to P. The first element, say $u_1 \in \mathcal{U}$, is chosen with probability p_{u_1}. Now we normalize the probabilities of the remaining elements by dividing with $1 - p_{u_1}$ and chose the next element, say u_2, with probability $p_{u_2}/(1 - p_{u_1})$ and so on until S contains L elements. The fact that different choosing sequences result in the same set S is taken into account by the sum over all permutations of [L].

Problem 5:
Establish an identification entropy for L-identification for sets which provides a lower bound for $\tilde{\mathcal{L}}^{L,q}(P^{(L)}, P)$?

We have seen that a crucial part in the discovery of the q-ary identification entropy of degree L and its role as a lower bound for L-identification is the Decomposition Lemma 4.10. We have

Problem 5.1:
Establish a decomposition formula for $\tilde{\mathcal{L}}^{L,q}(P^{(L)}, P)$ which is suitable to finding a solution for problem 5?

8 Appendix

```
LID {

  S_1 := [L];

  for i from 1 to ||c_v|| - 1 do {
    if (∀ l ∈ S_i : c_{u_l,i} ≠ c_{v,i}) then {
      return ("FALSE", i, ∅);
    }
    else {
      set S_{i+1} := {l ∈ S_i : c_{u_l,i} = c_{v,i}};
    }
  }

  if (∀ l ∈ S_{||c_v||} : c_{u_l,||c_v||} ≠ c_{v,||c_v||}) then {
    return ("FALSE", ||c_v||, ∅);
  }
  else {
    set S := {l ∈ S_{||c_v||} : c_{u_l,||c_v||} = c_{v,||c_v||}};
    return ("TRUE", ||c_v||, S);
  }
}
```

Table 8.1: The L-identification algorithm.

```
LIDforSets {

  S_1 := S

  for i from 1 to ||c_v|| do {

    if (∀ u ∈ S_i :  c_{u,i} ≠ c_{v,i}) then {
      return ("FALSE",i)
    }
    else {
      set S_{i+1} := {u ∈ S_i  :  c_{u,i} = c_{v,i}}
      if |S_{i+1}| = |N~(T_{c_v^i})| then {
        return ("TRUE",i)
      }
    }
  }
}
```

Table 8.2: The L-identification algorithm for sets.

List of Symbols

$2^{\mathcal{S}}$ ..... power set of S

$\mathcal{C}$ ..... a mapping from $\mathcal{U}$ to$\mathcal{Q}^*$, called a q-ary code on $\mathcal{U}$

$\mathcal{C}^n$ ..... the concatenated code of the basic code $\mathcal{C}$

$\mathcal{C}_{q,N}$ ..... the set of all q-ary balanced Huffman codes of size N

$\mathcal{C}_{q^n}$ ..... the q-ary code of size q^n all codewords having length n

c_u ..... codeword of $u \in \mathcal{U}$

c_u^k ..... prefix of length k of c_u

Δ_n ..... $\{(p_1, ..., p_n) \in [0,1]^n \ : \ 0 \leq \sum_{i=1}^n p_i \leq 1\}$

$\mathring{\Delta}_n$ ..... $\{(p_1, ..., p_n) \in (0,1)^n \ : \ 0 < \sum_{i=1}^n p_i \leq 1\}$

$\mathcal{L}^{L,q}(P)$ ..... $\min\limits_{\mathcal{C}} \mathcal{L}_{\mathcal{C}}^{L,q}(P)$, optimal worst-case (average) running time

$f_n \to a$ ..... a sequence tending to a as n goes to infinity

Γ_n ..... $\{(p_1, ..., p_n) \in [0,1]^n \ : \ \sum_{i=1}^n p_i = 1\}$

$\mathring{\Gamma}_n$ ..... $\{(p_1, ..., p_n) \in (0,1)^n \ : \ \sum_{i=1}^n p_i = 1\}$

$H_q(P)$ ..... Shannon's classical entropy for the alphabet size q

$H(P)$ ..... $H_2(P)$

$H_{\text{ID}}^{L,q}(P)$ ..... q-ary identification entropy of degree L

$\mathcal{H}_{q,N}$ ..... the set of all q-ary balanced Huffman trees with N leaves

$\mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v)$ .... L-identification running time for given u^L, v and q-ary code $\mathcal{C}$

$\mathcal{L}_{\mathcal{C}}^{L,q}(P, v)$ ..... $\sum\limits_{u^L \in \mathcal{U}^L} P_{u^L} \mathcal{L}_{\mathcal{C}}^{L,q}(u^L, v)$, average running time

$\mathcal{L}_{\mathcal{C}}^{L,q}(P)$ ..... $\max\limits_{v \in \mathcal{U}} \mathcal{L}_{\mathcal{C}}^{L,q}(P, v)$, worst-case (average) running time

$\mathcal{L}_{\mathcal{C}}^{L,q}(P,Q)$ .... $\sum_{v\in\mathcal{U}} Q(\{v\})\mathcal{L}_{\mathcal{C}}^{L,q}(P,v)$, expected (average) running time

$\mathcal{L}^{L,q}(P,Q)$ .... $\min_{\mathcal{C}} \mathcal{L}_{\mathcal{C}}^{L,q}(P,Q)$, optimal expected (average) running time

$\log_q$ ..... logarithm to the base q

$\log$ ..... $\log_2$

$[m+1,n]$ ..... $\{m+1,...,n\}$

$[n]$ ..... $\{1,2,...,n\}$

$\bar{\mathcal{N}}(T)$ ..... set of leaves of a tree T

$\mathring{\mathcal{N}}(T)$ ..... set inner nodes of a tree T

$\mathcal{N}(T)$ ..... $\bar{\mathcal{N}}(T) \cup \mathring{\mathcal{N}}(T)$

$\mathcal{Q}$ ..... $\{0,1,...,q-1\}$

q-bit ..... an element of $\mathcal{Q}$

$\mathcal{S}^*$ ..... $\bigcup_{d=0}^{\infty} \mathcal{S}^d$

$\mathcal{S}^c$ ..... complement of $\mathcal{S}$

$\binom{\mathcal{S}}{k}$ ..... the set of all k-element subsets of $\mathcal{S}$

$supp(P)$ ..... support of P

$T_{\mathcal{C}}$ ..... code tree of the code $\mathcal{C}$

T_x ..... the subtree of T with root in x

$\mathcal{U}$ ..... finite set, the output space

$(\mathcal{U},P)$ ..... source with output space $\mathcal{U}$ and output probability P

U ..... output random variable, $U = id_{\mathcal{U}}$

$(\mathcal{U}^n,P^n)$ ..... discrete memoryless source

$\mathcal{V}$ ..... finite set, the user space, w.l.o.g. $\mathcal{V}=\mathcal{U}$

V ..... user random variable, $V = id_{\mathcal{V}}$

$\underbrace{x,...,x}_{m}$ ..... block of m identical elements x

Bibliography

[1] J. Aczél and Z. Daróczy, "On Measures of Information and Their Characterizations", Mathematics in Science and Engineering, vol. 115, 1975.

[2] R. Ahlswede, "General theory of information transfer: updated", General Theory of Information Transfer and Combinatorics, Special issue of Discrete Applied Mathematics, to appear.

[3] R. Ahlswede, "Identification entropy", General Theory of Information Transfer and Combinatorics, Lecture Notes of Computer Science, vol. 4123, 2006.

[4] R. Ahlswede, B. Balkenhol and C. Kleinewächter, "Identification for sources", General Theory of Information Transfer and Combinatorics, Lecture Notes of Computer Science, vol. 4123, 2006.

[5] R. Ahlswede and N. Cai, "An interpretation of identification entropy", IEEE Trans. Inf. Theory, vol. 52, no. 9, 4198-4207, 2006.

[6] R. Ahlswede and G. Dueck, "Identification via channels", IEEE Trans. Inf. Theory, vol. 35, no.1, pp. 15-29, 1989.

[7] R. Ahlswede and G. Dueck, "Identification in the presence of feedback - a discovery of new capacity formulas", IEEE Trans. Inf. Theory, vol. 35, 30-39, 1989.

[8] I. Csiszár and J. Körner, "Information Theory: Coding Theorems for Discrete Memoryless Systems", Academic Press, 1981.

[9] T. M. Cover and J. A. Thomas, "Elements of Information Theory", Wiley-Interscience, 1991.

[10] R. M. Fano, "Transmission of Information", MIT Press, Cambridge MA and Wiley, 1961.

[11] T.S. Han and S. Verdú, "New results in the theory and application of identification via channels", IEEE Trans. on Inform. Theory, vol. IT-38, 14-25, 1992.

[12] D. A. Huffman, "A method for the construction of minimum-redundancy codes", Proceedings of the I.R.E., vol. 40, pp. 1098-1101, 1952.

[13] C. Kleinewächter, "On identification", General Theory of Information Transfer and Combinatorics, Lecture Notes of Computer Science, vol. 4123, 2006.

[14] J. G. Rosenstein, "Linear Orderings", Pure Appl. Math., vol. 98, Academic Press, 1982.

[15] C. E. Shannon, "A mathematical theory of communication", Bell System Tech. J., vol. 27, pp. 379-423 & 623-656, 1948.

[16] V. Strehl, private communication,Computer Science Department (Informatik 8), University Erlangen-Nürnberg, 2006.

[17] R. Veldhuis and M. Breeuwer, "An Introduction to Source Coding", Prentice Hall, 1993.